# HALAMAN JUDUL

# MODEL KLASIFIKASI ANALISIS DISKRIMINAN DENGAN METODE *PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS* (PLSDA) PADA DATA BERDIMENSI TINGGI

**SKRIPSI**

Sebagai salah satu syarat untuk memperoleh gelar Sarjana Statistika


Oleh:

**DURRAH IZZA ZHARFANI**

**155090507111009**



**PROGRAM STUDI SARJANA STATISTIKA**

**JURUSAN STATISTIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS BRAWIJAYA**

**MALANG**


**2019**

**LEMBAR PENGESAHAN SKRIPSI**

# MODEL KLASIFIKASI ANALISIS DISKRIMINAN DENGAN METODE *PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS* (PLSDA) PADA DATA BERDIMENSI TINGGI

Oleh:

**DURRAH IZZA ZHARFANI**
**155090507111009**

**Setelah dipertahankan di depan Majelis Penguji pada tanggal 22 April 2019 dan dinyatakan memenuhi syarat untuk memperoleh gelar Sarjana Statistika**

**Pembimbing,**

**Dr. Suci Astutik, S.Si., M.Si.**
**NIP. 197407221999032001**

**Mengetahui,**
**Ketua Jurusan Statistika**
**Fakultas MIPA Universitas Brawijaya**

**Rahma Fitriani, S.Si., M.Sc., Ph.D.**
**NIP. 197603281999032001**

## LEMBAR PERNYATAAN

**Nama : Durrah Izza Zharfani**
**NIM : 155090507111009**
**Jurusan : Statistika**
**Penulis Skripsi Berjudul :**

**MODEL KLASIFIKASI ANALISIS DISKRIMINAN DENGAN METODE *PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS* (PLSDA) PADA DATA BERDIMENSI TINGGI**

**Dengan ini menyatakan bahwa:**

1. **Isi dari skripsi yang saya buat adalah benar-benar karya sendiri dan tidak menjiplak karya orang lain, selain nama-nama yang termasuk di isi dan tertulis di daftar pustaka dalam Skripsi ini.**
2. **Apabila dikemudian hari ternyata Skripsi yang saya tulis terbukti hasil jiplakan, maka saya akan bersedia menanggung segala resiko yang akan saya terima.**

**Demikian pernyataan ini dibuat dengan segala kesadaran.**

**Malang, 2 Mei 2019**
**Yang menyatakan,**

**Durrah Izza Zharfani**
**NIM. 155090507111009**

# MODEL KLASIFIKASI ANALISIS DISKRIMINAN DENGAN METODE *PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS* (PLSDA) PADA DATA BERDIMENSI TINGGI

## ABSTRAK

Analisis diskriminan adalah analisis multivariat yang diterapkan untuk memodelkan hubungan antara peubah respon yang bersifat kategorik dengan peubah prediktor yang bersifat numerik. Analisis diskriminan yang memiliki jumlah peubah prediktor lebih banyak daripada jumlah amatannya harus ditangani dengan suatu metode yaitu metode *Partial Least Squares Discriminant Analysis* (PLSDA). Dampak adanya data berdimensi tinggi mengakibatkan prediksi yang dihasilkan kurang akurat. Salah satu contoh data berdimensi tinggi adalah data berupa gen-gen pada manusia yang menderita suatu jenis kanker. Tujuan penelitian ini adalah untuk mengetahui gen yang paling berpengaruh pada masing-masing jenis kanker. Pada penelitian ini digunakan data dengan jumlah gen sebanyak 685 dan amatan sebanyak 60 pasien yang diklasifikasikan ke dalam lima jenis kanker. Jenis kanker pada penelitian ini adalah BRCA (*Breast Carnicoma*), COAD (*Colon Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carnicoma*), LUAD (*Lung Adenocarcinoma*) dan PRAD (*Prostate Adenocarcinoma*). Hasil penelitian ini didapatkan gen-gen yang menjadi penciri pada masing-masing jenis kanker. Peubah penciri pada kanker jenis BRCA adalah Gen 452 dan Gen 634. Peubah penciri pada kanker jenis COAD adalah Gen 452 dan Gen 165. Peubah penciri pada kanker jenis KIRC adalah Gen 115 dan Gen 616. Peubah penciri pada kanker jenis LUAD adalah Gen 452 dan Gen 176. Peubah penciri pada kanker jenis PRAD adalah Gen 115 dan Gen 165.

**Kata Kunci**: *Data Dimensi Tinggi, PLSDA, Gen.*

# CLASSIFICATION MODEL OF DISCRIMINANT ANALYSIS WITH PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS (PLSDA) METHOD IN HIGH DIMENSION DATA


## ABSTRACT

Discriminant analysis is a multivariate analysis that is applied to model the relationship between categorical response variables with numerical predictor variables. Discriminant analysis that has more predictor variables than the number of observations must be handled by a method, such as Partial Least Squares Discriminant Analysis (PLSDA) method. The impact of the existence of high dimension data that the results in less accurate predictions. Example of high dimension data is data in the form of genes in human who suffer from a type of cancer. The purpose of this study was to determine the genes that most influence each type of cancer. In this study, the data used number of genes as much as 685 and the observations of 60 patients which classified into five types of cancer. The types of cancer in this study are BRCA (Breast Carnicoma), COAD (Colon Adenocarcinoma), KIRC (Kidney Renal Clear Cell Carnicoma), LUAD (Lung Adenocarcinoma) and PRAD (Prostate Adenocarcinoma). The results of this study obtained genes that become the characteristics in each type of cancer. Characteristics of cancer in BRCA are Gen 452 and Gen 634. Characteristics of cancer in COAD are Gen 452 and Gen 165. Characteristics of cancer in KIRC are Gen 115 and Gen 616. Characteristics of cancer in LUAD are Gen 452 and Gen 176. Characteristics of cancer in PRAD are Gen 115 and Gen 165.



**Keywords**: *High Dimension Data, PLSDA, Gen*

## KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan Skripsi dengan judul "Model Klasifikasi Analisis Diskriminan dengan Metode *Partial Least Squares Discriminant Analysis* (PLSDA) pada Data Berdimensi Tinggi".

Penulis menyadari bahwa penyusunan Skripsi ini tidak lepas dari bantuan dan dukungan oleh berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih dan rasa hormat kepada:

1. Ibu Dr. Suci Astutik, S.Si., M.Si. selaku Dosen Pembimbing Skripsi atas waktu, bimbingan dan nasihat yang senantiasa diberikan;
2. Ibu Dr. Dra. Ani Budi Astuti, M.Si. selaku Dosen Penguji I atas waktu, kritik dan saran yang telah diberikan;
3. Ibu Dr. Ir. Atiek Iriany, MS selaku Dosen Penguji II atas waktu, kritik dan saran yang telah diberikan;
4. Bapak Achmad Efendi, S.Si., M.Sc., Pd.D. selaku Ketua Program Studi Sarjana Statistika Fakultas MIPA Universitas Brawijaya;
5. Ibu Rahma Fitriani, S.Si., M.Sc., Ph.D. selaku Ketua Jurusan Statistika Fakultas MIPA Universitas Brawijaya;
6. Kedua orang tua, Ibu Dra. Sri Handayani dan Bapak Ir. Agung Budiono serta kakak tersayang Dioni Fadia Zatalini, S.Farm atas doa, didikan, dukungan dan kasih sayang yang senantiasa diberikan sepanjang hidup penulis;
7. Teman-teman Pondok Pesantren Darul Ulum, khususnya Muna Afdi Muniroh, Siti Nur Atiqoh, Alfiyya Dzakiyyarrohmah, Nuri Thobibatus Shofiya Al-Faruqi dan Nada Itorul Umam atas doa dan dukungan yang berarti bagi penulis;
8. Teman-teman Masta Argatyasa, khususnya Ilham Alifuddin Fitroini, Nadillah Nur Yasmin, Siti Nurlita Halisyah, Alayda Farah Dian, Tobias Surya, Iwan Gilang dan M. Adi Wibowo atas doa, dukungan, pengalaman, pelajaran dan kenangan indah;
9. Teman-teman Statistika UB 2015, khususnya Rizka Khaerani, Hafizh Iman Naufal, Alfan Alief Sektiananda dan Nawang Laksa Wiguna atas doa, dukungan, kebersamaan dan bantuan yang berarti bagi penulis selama kuliah di Universitas Brawijaya;
10. Teman-teman seperbimbingan Ibu Suci, khususnya Wahyu Romaningsih, Ivo Nila Krisna Putri, Novita Putri Kurnia Dewi,

Candra Indri Lestari, Annisa Nurvitadewi dan Ulva Nur Farida atas dukungan, doa, bantuan dan air mata selama pengerjaan Skripsi;

11. Teman-teman sepermainan, khususnya Wahyu Romaningsih, Ivo Nila Krisna Putri, Ravika Indiah Nirmala dan Ardiana Fatma Dewi atas doa, dukungan, bantuan dan kebersamaan selama hidup di Malang;
12. Seluruh jajaran dosen, staf dan karyawan Jurusan Statistika Universitas Brawijaya yang telah membantu proses penyelesaian Skripsi.

Penulis menyadari bahwa Skripsi ini masih jauh dari kata sempurna, maka dari itu penulis mengharapkan kritik dan saran yang membangun untuk penulisan lebih baik. Semoga Skripsi ini dapat memberikan manfaat dan pengetahuan bagi pembaca.

Malang, 2 Mei 2019

Penulis

# DAFTAR ISI

# DAFTAR TABEL

## DAFTAR GAMBAR

# DAFTAR LAMPIRAN

Hal.

# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Klasifikasi adalah pengelompokan objek-objek berdasarkan kesamaan sifat atau karakteristik. Tujuan klasifikasi adalah mengelompokkan suatu objek pada kelompok yang sudah ada sebelumnya. Dalam meneliti karakteristik data, ditentukan beberapa peubah penciri yang membedakan suatu kelompok dengan kelompok lainnya.

Pada bidang statistika dikenal banyak metode untuk mengklasifikasikan objek. Analisis diskriminan merupakan salah satu metode yang digunakan untuk mengklasifikasikan suatu objek ke dalam suatu kelompok yang sudah ditentukan sebelumnya. Analisis diskriminan juga digunakan untuk mengetahui peubah penciri yang membedakan anggota kelompok suatu populasi dan sebagai kriteria pengelompokan (Huberty, 1934). Analisis diskriminan dapat menggambarkan perbedaan antar kelompok populasi melalui suatu persamaan matematis yang disebut dengan fungsi diskriminan.

Data berdimensi tinggi merupakan data yang memiliki jumlah peubah prediktor lebih banyak daripada jumlah amatan. Semakin tinggi nilai dimensi, maka nilai informasi yang didapat semakin sulit diperoleh. Kejadian tersebut akan berdampak pada prediksi yang kurang akurat. Salah satu cara untuk menangani data dengan kasus tersebut melalui reduksi dimensi dengan metode pereduksian *Principal Component Analysis* (PCA) dan *Partial Least Squares* (PLS). Metode PLS memiliki keunggulan, yaitu tidak diperlukan asumsi, seperti normalitas dan multikolinieritas.

Pada dunia yang semakin maju ini, teknologi yang dibuat oleh manusia pun semakin canggih. Benda yang tidak dapat dilihat oleh mata manusia mampu dideteksi oleh alat yang canggih. Gen adalah unit pewarisan sifat bagi organisme hidup. Gen merupakan suatu benda yang sangat kecil dan tidak dapat dilihat dengan kasat mata. Terdapat banyak gen yang menyusun penyakit kanker. Adanya gen-gen tersebut mengklasifikasikan seseorang dikategorikan menderita suatu jenis kanker. Banyaknya gen tersebut, maka diperlukan analisis yang menangani kasus di mana peubah prediktor cukup banyak, sehingga dapat memprediksi data baru dengan tepat dan akurat.

Metode *Partial Least Squares Discriminant Analysis* (PLSDA) merupakan penggabungan teknik klasifikasi dengan *Partial Least Squares* (PLS) yang menghasilkan suatu prediksi klasifikasi yang lebih baik dan akurat. PLSDA akan mereduksi dimensi peubah awal dan menyusun model regresi secara simultan. PLSDA digunakan untuk mengklasifikasikan objek yang memiliki jumlah peubah prediktor lebih banyak daripada jumlah amatan. PLSDA sering digunakan pada data yang berdimensi tinggi dengan amatan kecil, misal pada data kemometrika. Penelitian sebelumnya dilakukan oleh Febbi Meidawati (2017) dengan judul Pemodelan Klasifikasi Obat Bahan Alam dengan Metode *Partial Least Squares Discriminant Analysis*. Penelitian tersebut menghasilkan nilai akurasi sebesar 86.67%. Berdasarkan nilai akurasi tersebut dapat dikatakan bahwa metode PLSDA merupakan metode yang tepat untuk mengklasifikasikan objek pada data berdimensi tinggi. Analisis diskriminan biasa kurang cocok jika digunakan untuk data dengan peubah prediktor yang cukup banyak karena model yang didapat akan sama dengan data contoh tetapi akan gagal dalam memprediksi data baru.

Pada penelitian ini, peneliti menggunakan metode *Partial Least Squares Discriminant Analysis* (PLSDA) untuk memodelkan klasifikasi penyakit kanker berdasarkan gen yang ada pada manusia dan untuk mengetahui gen apa saja yang menjadi pembeda pada masing-masing kanker.

## 1.2. Rumusan Masalah

Rumusan masalah pada penelitian ini adalah:

1) Bagaimana fungsi diskriminan yang terbentuk untuk mengklasifikasikan pasien ke dalam lima jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD)?
2) Seberapa tepat hasil klasifikasi pada fungsi diskriminan yang terbentuk?
3) Peubah prediktor apa yang paling berpengaruh dalam menjelaskan klasifikasi pasien ke dalam lima jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD)?

## 1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1) Membentuk fungsi diskriminan yang terbentuk untuk mengklasifikasikan pasien ke dalam lima jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD).
2) Memperoleh tingkat ketepatan model klasifikasi pada fungsi diskriminan yang terbentuk.
3) Mengidentifikasi peubah prediktor yang paling berpengaruh dalam menjelaskan klasifikasi pasien ke dalam lima jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD).

## 1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah model diharapkan mampu mengklasifikasikan objek ketika jumlah peubah prediktor cukup banyak sehingga didapatkan prediksi yang tepat dan akurat. Bagi dunia medis, manfaat adanya penelitian ini adalah mempermudah dalam menentukan klasifikasi seseorang terkena kanker.

## 1.5. Batasan Masalah

Batasan masalah pada penelitian ini adalah kanker yang digunakan hanya lima jenis, yaitu BRCA (*Breast Carnicoma*), COAD (*Colon Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carnicoma*), LUAD (*Lung Adenocarcinoma*) dan PRAD (*Prostate Adenocarcinoma*). Banyaknya gen yang digunakan pada penelitian ini hanya melibatkan 685 gen pertama dengan amatan sebanyak 60 pasien yang tercatat di Machine Learning Repository.

# BAB II
# TINJAUAN PUSTAKA

## 2.1. Analisis Diskriminan

Analisis multivariat diklasifikasikan menjadi dua, yaitu teknik dependensi dan teknik interdependensi. Teknik dependensi adalah teknik statistika multivariat yang digunakan untuk menguji hubungan antar peubah di mana peubah respon dan peubah prediktor sudah dapat dibedakan (Mattjik dan Sumertajaya, 2011). Analisis diskriminan merupakan analisis yang termasuk dalam teknik dependensi.

Menurut Hair dkk. (1998), analisis diskriminan adalah analisis multivariat yang diterapkan untuk memodelkan hubungan antara satu peubah respon yang bersifat dikotom atau multikotom dan merupakan data non-metrik (nominal dan ordinal) dengan peubah prediktor yang bersifat kuantitatif. Menurut Johnson dan Wichern (2007), analisis diskriminan adalah suatu teknik peubah ganda yang digunakan untuk memisahkan pengamatan atau objek ke dalam kelompok atau himpunan yang berbeda dan untuk mengklasifikasikan objek baru ke dalam salah satu kelompok yang telah ditentukan sebelumnya. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu individu atau observasi ke dalam kelompok yang saling bebas (*mutually exclusive/disjoint*) dan menyeluruh berdasarkan sejumlah peubah prediktor (Mattjik dan Sumertajaya, 2011).

Berdasarkan jumlah klasifikasi pada peubah respon, analisis diskriminan dibagi menjadi dua, yaitu *Two Group Discriminant Analysis* dan *Multiple Discriminant Analysis* (Hair dkk., 2010). Jika terdapat dua klasifikasi peubah respon maka analisis yang digunakan adalah *Two Group Discriminant Analysis*. Jika terdapat tiga atau lebih klasifikasi peubah respon maka analisis yang digunakan adalah *Multiple Discriminant Analysis*.

Model fungsi analisis diskriminan adalah sebuah persamaan yang menunjukkan suatu kombinasi linier dari berbagai peubah prediktor. Menurut Hair (2010), fungsi diskriminan ditunjukkan pada persamaan (2.1).

$$D_g = a + W_1 X_{1g} + W_2 X_{2g} + W_3 X_{3g} + \cdots + b_p X_{pg} \qquad (2.1)$$

Keterangan:

$D$ : skor diskriminan

$a$ : intersep

$W$ : koefisien diskriminan atau bobot
$X$ : peubah prediktor
$p$ : banyaknya peubah prediktor
$g$ : banyaknya kelompok

## 2.2. Data Hilang (*Missing Value*)

Data hilang (*missing value*) adalah suatu kondisi hilangnya sebagian fitur pada data set. *Missing value* terjadi karena informasi tentang objek tidak diberikan, sulit dicari atau informasi tersebut tidak tersedia. *Missing value* juga dapat disebabkan oleh kesalahan sistem seperti tidak adanya respon terhadap sensor atau perangkat penerima *input*. Dapat pula disebabkan oleh *human error* seperti ketidaklengkapan memasukkan data pada *database*. *Missing value* sering terjadi pada *data mining* dan data berdimensi tinggi karena jumlah data yang cukup banyak.

Banyaknya *missing value* pada suatu data tidak akan bermasalah apabila terdapat sekitar 1% dari keseluruhan data sehingga *missing value* tersebut dapat diabaikan. Apabila jumlah data hilang cukup besar maka perlu dilakukan penanganan. Menurut Han dkk. (2012), penanganan *missing value* pada suatu data dapat dilakukan dengan dua cara. Pertama, memasukkan rata-rata atau median dari peubah yang mengandung *missing value*. Kedua, menyisihkan peubah yang mengandung *missing value*.

## 2.3. Analisis Korelasi

Analisis korelasi merupakan suatu analisis untuk mengetahui kekuatan hubungan antara dua peubah melalui sebuah bilangan yang disebut koefisien korelasi (Walpole, 1993). Koefisien korelasi antara dua peubah adalah suatu ukuran hubungan linier antara kedua peubah tersebut. Koefisien korelasi memiliki rentang nilai -1 sampai 1. Apabila nilai koefisien korelasi bernilai 0, maka dapat dikatakan bahwa kedua peubah tidak memiliki hubungan. Apabila nilai koefisien korelasi bernilai -1 atau 1, maka dapat dikatakan bahwa kedua peubah memiliki hubungan yang erat.

Analisis korelasi dapat dilakukan dengan beberapa pengujian tergantung pada jenis data. Uji korelasi Pearson merupakan pengujian korelasi ketika jenis data bersifat kuantitatif (data berskala interval atau rasio). Dalam pengujian akan menghasilkan koefisien korelasi yang digunakan untuk menyatakan besar hubungan linier antara dua

peubah. Menurut Walpole (1993), rumus untuk menghitung koefisien korelasi Pearson ditunjukkan pada persamaan (2.2).

$$r = \frac{\sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sqrt{\sum X_1^2 - \frac{(\sum X_1)^2}{n}} \sqrt{\sum X_2^2 - \frac{(\sum X_2)^2}{n}}} \quad (2.2)$$

Keterangan:
$r$ : koefisien korelasi
$n$ : banyaknya amatan

Statistik uji yang digunakan untuk mengetahui tingkat signifikansi dari koefisien korelasi adalah statistik uji t dengan rumus sesuai persamaan (2.3).

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.3)$$

Hipotesis untuk pengujian korelasi sebagai berikut.
$H_0 : \rho = 0$ vs
$H_1 : \rho \neq 0$

Keputusan untuk menolak $H_0$ apabila $|t_{hitung}| > t_{(\frac{\alpha}{2}, n-2)}$ atau dengan melihat *p-value*$< \alpha = 0{,}05$, artinya terdapat hubungan yang signifikan antara dua peubah. Sebaliknya, keputusan untuk meneriman $H_0$ apabila $|t_{hitung}| \leq t_{(\frac{\alpha}{2}, n-2)}$ atau dengan melihat *p-value* $\geq \alpha = 0{,}05$, artinya tidak terdapat hubungan yang signifikan antara dua peubah.

## 2.4. *Partial Least Squares Discriminant Analysis* (PLSDA)

*Partial Least Squares* (PLS) merupakan metode yang pertama kali dikembangkan oleh Herman Wold pada tahun 1966 pada bidang ekonometrika. PLS merupakan teknik analisis yang *powerfull* karena dapat diterapkan pada semua data dan ukuran sampel tidak harus besar (Jaya dan Sumertajaya, 2008). PLS juga mampu menangani banyak peubah prediktor, bahkan sekalipun terjadi multikolinieritas di antara peubah-peubah tersebut (Ramzan dan Khan, 2010). Menurut Wold dkk. (2001), PLS dapat digunakan untuk menganalisis data yang memiliki korelasi tinggi, *noise* (derau) dan memiliki banyak peubah

prediktor serta dapat memodelkan secara simultan beberapa peubah respon.

*Partial Least Squares Regression* (PLSR) digunakan untuk menangani adanya korelasi antar peubah prediktor menjadi komponen baru dengan peubah yang tidak berkorelasi (Cinca dan Nieto, 2011). *Partial Least Squares Discriminant Analysis* (PLSDA) merupakan regresi PLS klasik yang memiliki peubah respon bersifat kategorik yang menunjukkan suatu kelas klasifikasi. Metode ini merupakan gabungan antara analisis diskriminan biasa dan analisis diskriminan pada komponen utama yang penting dari peubah prediktor (Enciso dan Tenenhaus, 2003). PLSDA akan mereduksi dimensi peubah awal dan menyusun model regresi secara simultan.

Regresi PLS menguraikan peubah **X** (matriks berukuran $n \times p$) dan peubah **Y** (matriks berukuran $n \times g$) sebagai sekumpulan nilai ortogonal dan sekumpulan nilai yang khusus dengan membentuk komponen PLS. Persamaan regresi PLS dijelaskan pada persamaan (2.4) dan diduga dengan $\boldsymbol{\beta}$ (matriks berukuran $p \times g$) sesuai persamaan (2.5).

$$\boldsymbol{Y} = \boldsymbol{X\beta} \tag{2.4}$$

dengan

$$\boldsymbol{\beta} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{C}' \tag{2.5}$$

Dalam menentukan skor X (**T**) terlebih dahulu mencari nilai bobot X (**W**) dan penentuan skor Y (**U**) terlebih dahulu mencari nilai bobot Y (**C**). Nilai **W** dan **C** dapat diperoleh sesuai pada persamaan (2.6) dan persamaan (2.7).

$$\boldsymbol{T} = \boldsymbol{XW} \tag{2.6}$$

$$\boldsymbol{U} = \boldsymbol{YC} \tag{2.7}$$

Keterangan:
$\boldsymbol{T}$ : skor X (matriks berukuran $n \times a$)
$\boldsymbol{U}$ : skor Y (matriks berukuran $n \times a$)
$\boldsymbol{W}$ : bobot X (matriks berukuran $p \times a$)
$\boldsymbol{C}$ : bobot Y (matriks berukuran $g \times a$)
$\boldsymbol{P}$ : X-*loadings* (matriks berukuran $p \times a$)
$g$ : banyaknya kelompok

$p$ : banyaknya peubah prediktor
$n$ : banyaknya amatan
$a$ : banyaknya komponen

Algoritma standar untuk menghitung komponen PLS adalah *Nonlinear Iterative Partial Least Squares* (NIPALS). Ide dasar dalam algoritma ini adalah mengestimasi parameter **t** dan **u** dengan suatu proses iteratif dari regresi *least square*. Langkah-langkah dalam algoritma NIPALS dijelaskan pada persamaan (2.8) sampai persamaan (2.21).

1) Inisialisasi skor Y pertama dari salah satu kolom pada data Y.

$$\boldsymbol{u_f} = \boldsymbol{y_k}\ f = 1, 2, \dots, a; k = 1, 2, \dots, g \tag{2.8}$$

2) Hitung bobot X.

$$\boldsymbol{w_f} = \frac{(\boldsymbol{X'u_f})}{(\boldsymbol{u_f}'\boldsymbol{u_f})} \tag{2.9}$$

3) Bakukan nilai *w*.

$$\boldsymbol{w_f} = \frac{\boldsymbol{w_f}}{\|\boldsymbol{w_f}\|} \tag{2.10}$$

4) Hitung skor X.

$$\boldsymbol{t_f} = \boldsymbol{X w_f} \tag{2.11}$$

5) Hitung bobot Y.

$$\boldsymbol{c_f} = \frac{(\boldsymbol{Y't_f})}{(\boldsymbol{t_f}'\boldsymbol{t_f})} \tag{2.12}$$

6) Bakukan nilai *c*.

$$\boldsymbol{c_f} = \frac{\boldsymbol{c_f}}{\|\boldsymbol{c_f}\|} \tag{2.13}$$

7) Hitung skor Y baru.

$$\boldsymbol{u_f}^* = \boldsymbol{Y c_f} \tag{2.14}$$

8) Tentukan selisih skor Y.

$$\boldsymbol{u_\Delta} = \boldsymbol{u_f}^* - \boldsymbol{u_f} \tag{2.15}$$

$$\Delta \boldsymbol{u} = \boldsymbol{u}_{\Delta}{}' \boldsymbol{u}_{\Delta} \quad (2.16)$$

9) Jika $\Delta \boldsymbol{u} > \varepsilon$ maka lakukan kembali sesuai langkah 2 dengan menggunakan $\boldsymbol{u}_{\boldsymbol{f}}{}^{*}$. Jika $\Delta \boldsymbol{u} < \varepsilon$ maka didapatkan komponen PLS pertama dan dilanjutkan ke langkah 10.

10) Hitung X-*loadings*.

$$\boldsymbol{p}_{\boldsymbol{f}} = \frac{(\boldsymbol{X}' \boldsymbol{t}_{\boldsymbol{f}})}{(\boldsymbol{t}_{\boldsymbol{f}}{}' \boldsymbol{t}_{\boldsymbol{f}})} \quad (2.17)$$

11) Hitung Y-*loadings*.

$$\boldsymbol{q}_{\boldsymbol{f}} = \frac{(\boldsymbol{u}_{\boldsymbol{f}}{}' \boldsymbol{t}_{\boldsymbol{f}})}{(\boldsymbol{t}_{\boldsymbol{f}}{}' \boldsymbol{t}_{\boldsymbol{f}})} \quad (2.18)$$

12) Hitung sisaan dari X.

$$\boldsymbol{X}_{\boldsymbol{res}} = \boldsymbol{X} - \boldsymbol{t}_{\boldsymbol{f}} \boldsymbol{p}_{\boldsymbol{f}}{}' \quad (2.19)$$

13) Tentukan nilai *Y* baru dengan estimasi

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{baru}} = \widehat{\boldsymbol{Y}}_{\boldsymbol{inisial}} - \boldsymbol{t}_{\boldsymbol{f}} \boldsymbol{q}_{\boldsymbol{f}}{}' \quad (2.20)$$

14) Hitung sisaan dari *Y*

$$\boldsymbol{Y}_{\boldsymbol{res}} = \boldsymbol{Y}_{\boldsymbol{awal}} - \widehat{\boldsymbol{Y}}_{\boldsymbol{baru}} \quad (2.21)$$

15) Ganti nilai *X* dengan $X_{res}$ dan nilai *Y* dengan $Y_{res}$, kemudian lakukan iterasi lagi mulai dari tahap 2 untuk mendapatkan komponen PLS lainnya.

## 2.5. Validasi Silang *(Cross Validation)*

Kohavi (1995) menyatakan bahwa terdapat beberapa metode untuk menguji keakuratan suatu model klasifikasi, antara lain: *holdout, cross validation* dan *bootstrap*. *Cross validation* atau validasi silang adalah metode statistik yang digunakan untuk mengevaluasi kinerja model atau algoritma yang membagi data menjadi data *training* dan data *testing*. Data *training* adalah data yang digunakan untuk membentuk model sedangkan data *testing* adalah data independen yang tidak terlibat dalam pembentukan model dan digunakan untuk menguji keakuratan model dalam menduga data baru. Selain memberikan informasi mengenai nilai keakuratan,

validasi silang juga mampu memberikan informasi mengenai banyaknya amatan yang tepat dan tidak tepat diklasifikasikan. Metode ini digunakan untuk mengatasi *overfitting*. *Overfitting* adalah kondisi ketika model sudah sesuai dengan data contoh tetapi kurang baik dalam memprediksi data yang bukan bagian dari data penyusun model. Hal tersebut dapat terjadi apabila jumlah peubah prediktor lebih banyak dibanding dengan jumlah amatan.

Salah satu bentuk validasi silang adalah *k-fold cross validation* yang artinya terdapat sebanyak *k* pengujian. *K-fold cross validation* adalah metode pengujian yang diterapkan pada data *training* untuk menguji tingkat validasi (termasuk besaran *error*) dari data yang diuji. Metode *k-fold cross validation* membagi data menjadi *k* bagian yang sama secara acak. Satu bagian menjadi data *testing* dan *k-1* bagian menjadi data *training*. Tabel 2.1 merupakan ilustrasi penerapan data *testing* dan data *training* pada *cross validation* sebanyak 5-*fold*.

Tabel 2.1. Ilustrasi Data *Testing* dan Data *Training* pada 5-*Fold Cross Validation*

| | | | | | |
|---|---|---|---|---|---|
| *Fold* 1 | *testing* | *training* | *training* | *training* | *training* |
| *Fold* 2 | *training* | *testing* | *training* | *training* | *training* |
| *Fold* 3 | *training* | *training* | *testing* | *training* | *training* |
| *Fold* 4 | *training* | *training* | *training* | *testing* | *training* |
| *Fold* 5 | *training* | *training* | *training* | *training* | *testing* |

Setiap *fold* memiliki kesempatan satu kali menjadi data *testing* dan empat kali menjadi data *training*. Posisi data *testing* harus selalu berbeda di setiap *fold*. Misal di *fold* 1 data *testing* berada di posisi pertama, kemudian *fold* 2 berada di posisi kedua, dan seterusnya. Setiap validasi silang terdiri dari 4-*fold* yang bertindak sebagai data *training* dan 1-*fold* yang bertindak sebagai data *testing*.

Rumus untuk perhitungan nilai akurasi dengan metode *cross validation* dijelaskan pada persamaan (2.22).

$$acc_{CV} = \frac{1}{n}\sum_{i=1}^{k}\left(\frac{C}{D_{(k)}}\right)_i \tag{2.22}$$

Keterangan:
$C$ : total amatan yang tepat diklasifikasikan
$D_{(k)}$ : banyaknya amatan pada *fold k*

## 2.6. Ketepatan Model Klasifikasi

Salah satu metode untuk mengukur ketepatan model klasifikasi adalah dengan *confusion matrix* atau dikenal dengan istilah tabel kontingensi. Tabel kontingensi mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan model dengan hasil klasifikasi yang sebenarnya. Tabel 2.2 merupakan contoh tabel kontingensi yang memiliki dua kategori kelompok.

Tabel 2.2. Tabel Kontingensi dengan Dua Kategori Kelompok

| Prediksi | Aktual | |
|---|---|---|
| | Positif | Negatif |
| Positif | *True Positive* (TP) | *False Positive* (FP) |
| Negatif | *False Negative* (FN) | *True Negative* (TN) |

Nilai *True Positive* (TP) adalah data positif yang terdeteksi benar. Sama halnya dengan TP, *True Negative* (TN) untuk data negatif yang terdeteksi benar. Nilai *False Positive* (FP) yaitu data negatif yang terdeteksi sebagai data positif. Sebaliknya, nilai *False Negative* (FN) untuk data positif yang terdeteksi sebagai data negatif. Semakin banyak amatan yang terdeteksi benar maka nilai akurasinya semakin tinggi. Nilai akurasi berada pada rentang nilai 0 sampai 1 dengan nilai 1 merupakan nilai dengan akurasi tertinggi.

## 2.7. *Variable Importance in Projection* (VIP)

Skor *VIP* memberikan arti pengaruh setiap peubah prediktor terhadap model. Menurut Farres dkk. (2015), skor VIP merupakan jumlah kuadrat terboboti dari nilai *weight* PLS ($w$) yang dihitung dari total keragaman $Y$ yang dijelaskan oleh setiap peubah laten. Peubah laten adalah peubah yang tidak dapat diukur secara langsung melainkan diukur secara tidak langsung dengan bantuan beberapa indikator. Skor VIP digunakan sebagai acuan untuk memilih peubah prediktor yang paling berpengaruh dalam menjelaskan peubah respon. Rumus VIP dijelaskan pada persamaan (2.23).

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \times SSY_f \times p}{SSY_{total} \times a}} \quad (2.23)$$

Keterangan:

$w_{jf}$ : nilai *weight* dari peubah prediktor ke-$j$ dan komponen ke-$f$

$SSY_f$ : jumlah kuadrat keragaman komponen ke-$f$

$p$ : jumlah peubah prediktor

$SSY_{total}$ : jumlah kuadrat total keragaman peubah respon

$a$ : jumlah komponen yang terbentuk

$VIP_j$ merupakan ukuran kontribusi masing-masing peubah prediktor sesuai dengan varian yang dijelaskan oleh masing-masing komponen PLS. Suatu peubah prediktor dikatakan berpengaruh dalam menjelaskan peubah respon apabila skor VIP > 1.

## 2.8. Gen

Istilah gen diciptakan oleh W. Johannsen pada tahun 1909. Gen adalah unit pewarisan sifat bagi organisme hidup. Gen terdiri dari DNA yang diselubungi dan diikat oleh protein. Secara kimia, dapat disebut bahwa unit informasi genetik adalah DNA. Ukuran gen ditaksir 4 sampai 50 mμ. Bentuk fisik gen adalah urutan DNA yang melekat atau berada di suatu protein, polipeptida, atau seuntai RNA yang memiliki fungsi bagi organisme yang memilikinya. Pada molekul DNA terdapat gen, dalam hal ini gen merupakan urutan nukleotida tertentu dari DNA yang mengekspresikan sifat tertentu yang mengkode pembentukan suatu polipeptida, yang mengkode pembentukan suatu RNA atau yang dibutuhkan untuk transkripsi gen lain.

Ekspresi gen adalah tingkat paling mendasar di mana genotip pada suatu individu memunculkan fenotip, yaitu sifat yang dapat diamati. Adanya RNA *microarrays* dan RNA *sequence* (untaian RNA) membantu mempelajari transkriptom yang merupakan seperangkat transkrip RNA lengkap yang diproduksi oleh genom dalam keadaan tertentu dalam sel atau jaringan tertentu (Stefanska dan MacEwan, 2017). Perbandingan transkriptom memungkinkan identifikasi gen yang diekspresikan secara berbeda dalam populasi sel yang berbeda dan akibatnya membantu dalam menjelaskan mekanisme penyakit, terutama pada kanker.

## 2.9. Kanker

Tumor adalah pertumbuhan sel-sel tubuh yang abnormal. Sel merupakan unit terkecil yang menyusun jaringan tubuh manusia. Masing-masing sel mengandung gen yang berfungsi untuk menentukan pertumbuhan, perkembangan, atau perbaikan yang terjadi dalam tubuh. Pembelahan sel dikontrol oleh suatu kontrol genetik. Gen-gen tertentu harus mengatur proses pembelahan sel. Gen-gen pengatur ini seperti halnya dengan gen-gen lain, juga dapat mengalami mutasi. Mutasi yang menghilangkan fungsi dari gen-gen pengatur ini dapat mengantar terjadinya pembelahan sel secara abnormal. Pertumbuhan dan pembelahan sel yang tidak terkontrol dan menghasilkan suatu massa sel-sel disebut dengan tumor (Suryo, 1990). Tumor yang ganas akan melepaskan sel-sel dan akan dibawa dengan aliran darah ke bagian lain dari tubuh, disebut dengan kanker. Proses ini dinamakan metastase. Tumor yang tidak ganas tidak mengalami metastase. Berikut merupakan penjelasan mengenai beberapa jenis kanker menurut National Cancer Institute (NCI).

1) BRCA (*Breast Carcinoma*)

   BRCA atau *breast cancer* adalah kanker yang berkembang dari jaringan payudara. Kanker payudara dimulai ketika sel-sel di payudara mulai tumbuh di luar kendali. Sel-sel ini biasanya membentuk tumor yang sering terlihat pada *rontgen* atau terasa sebagai benjolan.

2) COAD (*Colon Adenocarcinoma*)

   COAD merupakan salah satu jenis kanker ganas yang terjadi pada epitel mukosa usus besar dari kolon sampai dengan rektum.

3) KIRC (*Kidney Renal Clear Cell Carcinoma*)

   KIRC adalah jenis kanker ginjal yang paling umum. Kanker ini terbentuk di sel-sel yang melapisis tubulus kecil di ginjal yang menyaring limbah dari darah dan membuat urin.

4) LUAD (*Lung Adenocarcinoma*)

   LUAD adalah jenis kanker paru yang paling umum dan lebih banyak muncul pada wanita. Kanker paru adalah pertumbuhan sel kanker yang tidak terkendali dalam jaringan paru. Kanker paru disebabkan oleh sejumlah karsinogen, terutama asap rokok. Faktor lain seseorang terkena kanker paru adalah polusi udara, paparan radon, genetik, dan lingkungan. Kanker paru jenis LUAD berkembang dari sel-sel yang memproduksi lendir pada permukaan saluran udara.

5) PRAD (*Prostate Adenocarcinoma*)

Prostat merupakan kelenjar yang hanya ditemukan pada organ tubuh pria. Sebagian besar sel dalam kelenjar prostat adalah jenis kelenjar, yang berarti bahwa adenokarsinoma adalah jenis kanker yang paling umum terjadi di prostat. Sekitar 5-10% kanker prostat disebabkan oleh cacat genetik, sehingga pria yang memiliki riwayat keluarga lebih berisiko mengembangkannya sendiri. Pria yang memiliki kerabat dekat (saudara, ayah, anak) yang menderita kanker prostat dua atau tiga kali lebih mungkin mengembangkan kanker prostat sendiri.

# BAB III
# METODE PENELITIAN

## 3.1. Data

Data yang digunakan pada penelitian ini adalah data sekunder yang didapatkan dari Machine Learning Repository pada *link* https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq. Data berupa panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker yang dilambangkan dengan dummy (gene_XX). Terdapat lima jenis kanker, yaitu BRCA, KIRC, COAD, LUAD, dan PRAD. Tingkat ekspresi gen pasien bertindak sebagai peubah prediktor sedangkan klasifikasi jenis kanker bertindak sebagai peubah respon. Peubah prediktor yang digunakan sebanyak 685 gen dengan sampel sebanyak 60 amatan. Gambaran umum data dapat dilihat pada Lampiran 1. Untuk mempermudah pekerjaan, jenis kanker dilambangkan dengan angka seperti pada Tabel 3.1.

Tabel 3.1. Kategori Jenis Kanker

| Jenis Kanker | Kategori |
|---|---|
| BRCA | 1 |
| COAD | 2 |
| KIRC | 3 |
| LUAD | 4 |
| PRAD | 5 |

Dalam menganalisis data menggunakan metode PLSDA, sebelumnya data tersebut dibagi menjadi dua, yaitu data *training* dan data *testing*. Data *training* digunakan untuk pemodelan klasifikasi sedangkan data *testing* digunakan untuk pengujian pada data baru. Penelitian ini menggunakan persentase 80% untuk data *training* dan 20% untuk data *testing*.

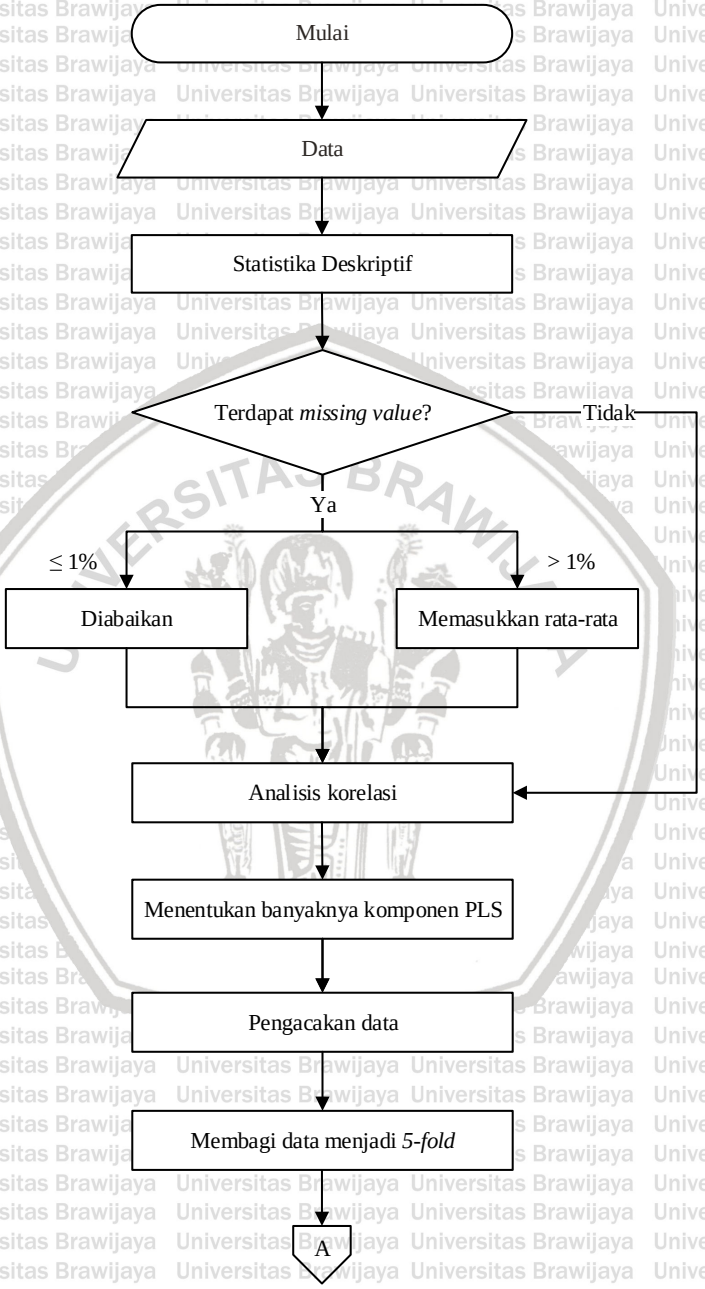
## 3.2. Metode Analisis Data

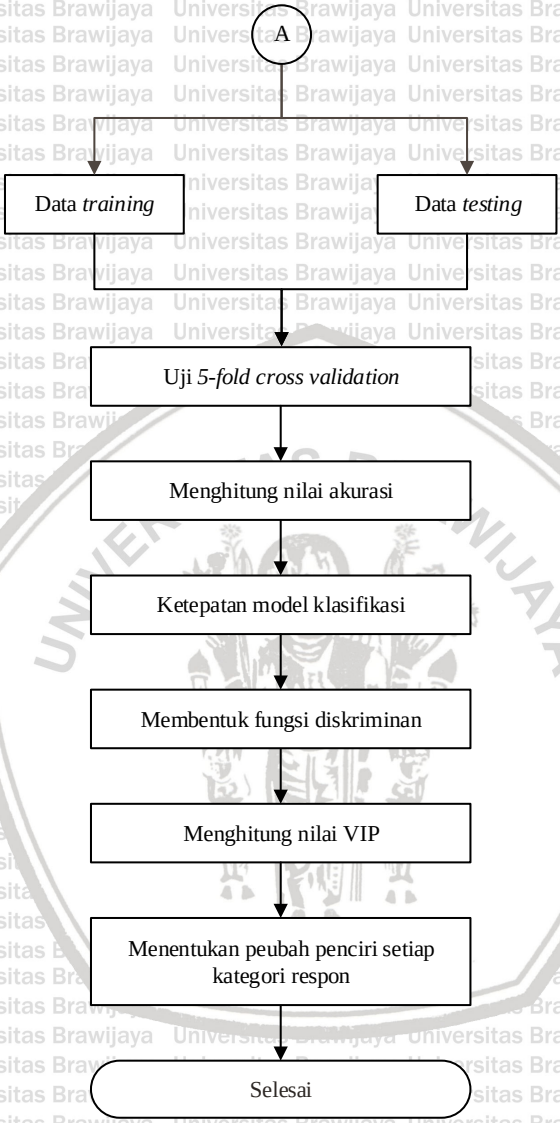
Dalam penelitian ini, analisis yang digunakan adalah analisis diskriminan dengan metode *Partial Least Squares Discriminant Analysis* (PLSDA) menggunakan bantuan dua *software*, yaitu Minitab

dan R Studio dengan *packages* DiscriMiner. Langkah-langkah dalam melakukan analisis ini dijelaskan sebagai berikut.

1) Analisis statistika deskriptif.
2) Mengidentifikasi adanya data hilang (*missing value*) sesuai pada sub bab 2.2.
3) Mengetahui hubungan antar peubah prediktor dengan analisis korelasi sesuai pada sub bab 2.3.
4) Menentukan komponen PLS yang digunakan dalam model dengan melihat proporsi keragaman peubah $X$ dan peubah $Y$ beserta kumulatifnya pada masing-masing komponen. Pemilihan banyaknya komponen berdasarkan keragaman kumulatif dari peubah $X$ dan peubah $Y$ yang mencapai 80%.
5) Mengacak data.
6) Melakukan uji *k-fold cross validation* dengan $k$=5. Berikut tahapan pengujian *cross validation*.
   a. Data dibagi menjadi 5-*fold* berukuran sama.
   b. Melakukan validasi silang sebanyak lima kali. Setiap validasi silang terdiri dari 4-*fold* yang bertindak sebagai data *training* dan 1-*fold* bertindak sebagai data *testing*. Setiap *fold* terdiri dari 1 data *testing* dan 4 data *training* seperti pada Tabel 2.2.
   c. Untuk data *training* dibentuk model klasifikasi dengan metode PLSDA sedangkan data *testing* dilakukan untuk pengujian. Hasil dari data *training* dan data *testing* ditampilkan pada tabel kontingensi sebagaimana ditunjukkan pada Tabel 2.3.
   d. Menghitung nilai akurasi dari keseluruhan validasi silang dengan persamaan (2.21).
7) Mengidentifikasi amatan-amatan yang tidak tepat diklasifikasikan oleh model sesuai pada sub bab 2.6.
8) Membentuk fungsi diskriminan sesuai pada persamaan (2.1).
9) Mengidentifikasi peubah prediktor penting dalam model dengan melihat nilai VIP sesuai persamaan (2.22).
10) Mengidentifikasi peubah penciri pada setiap kategori respon.

Langkah-langkah penelitian secara lengkap disajikan pada Gambar 3.1.


Mulai
Data
Statistika Deskriptif
Terdapat *missing value*?
Tidak
Ya
≤ 1%
> 1%
Diabaikan
Memasukkan rata-rata
Analisis korelasi
Menentukan banyaknya komponen PLS
Pengacakan data
Membagi data menjadi *5-fold*
A

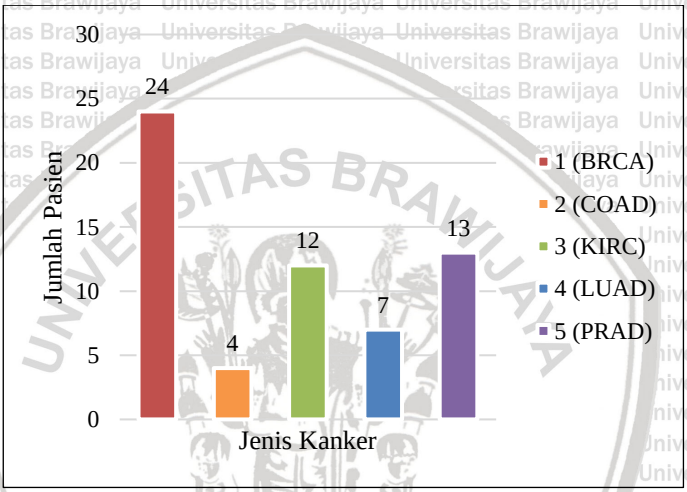



Gambar 3.1. Diagram Alir Penelitian

# BAB IV
# HASIL DAN PEMBAHASAN

## 4.1. Statistika Deskriptif

Sebelum dilakukan analisis diskriminan dengan metode PLSDA, terlebih dahulu dilakukan statistika deskriptif untuk mengetahui gambaran data. Gambar 4.1 merupakan diagram data gen pasien yang menderita lima jenis kanker.




Gambar 4.1. Statistika Deskriptif

Berdasarkan Gambar 4.1, dapat dilihat bahwa dari 60 pasien penderita kanker, persentase tertinggi berada pada kanker jenis BRCA yaitu sebesar 40% (24 pasien). Sedangkan persentase terendah berada pada kanker jenis COAD yaitu sebesar 7% (4 pasien). Tiga jenis kanker lain, yaitu KIRC, LUAD dan PRAD masing-masing memiliki persentase 20% (12 pasien), 11% (7 pasien) dan 22% (13 pasien).

## 4.2. Identifikasi Data Hilang (*Missing Value*)

Data gen pasien penderita kanker merupakan data berdimensi tinggi, artinya jumlah data cukup banyak. Perlu diidentifikasi adanya *missing value* pada data. Apabila jumlah *missing value* sekitar 1% dari keseluruhan data maka *missing value* tersebut dapat diabaikan. Apabila jumlah *missing value* cukup besar maka perlu dilakukan penanganan.

*Missing value* dapat diidentifikasi dengan analisis statistika deskriptif menggunakan *software* Minitab. Secara keseluruhan, data gen pasien yang menderita kanker terdapat 41.100 data. Setelah dilakukan analisis statistika deskriptif, terdapat 427 dari 41.100 data *missing value*. Artinya, terdapat 0,01% *missing value* pada data gen pasien penderita kanker. Persentase data hilang tersebut cukup kecil sehingga data hilang diabaikan. Banyaknya data hilang pada masing-masing peubah prediktor dapat dilihat di Lampiran 2.

### 4.3. Identifikasi Hubungan Antar Peubah Prediktor

Analisis korelasi yang digunakan adalah uji korelasi Pearson dengan hipotesis sebagai berikut.

$H_0 : \rho = 0$ vs

$H_1 : \rho \neq 0$

Peubah prediktor sebanyak 685 dilakukan uji korelasi antar dua peubah sehingga terdapat 234.270 pengujian korelasi. Sebanyak 196.400 pengujian menghasilkan keputusan terima $H_0$, artinya sebanyak 196.400 pengujian tidak terdapat hubungan yang signifikan antar dua peubah prediktor. Sebanyak 37.870 pengujian menghasilkan keputusan tolak $H_0$, artinya sebanyak 37.870 pengujian terdapat hubungan yang signifikan antar dua peubah prediktor. Kesimpulan yang didapatkan dari pengujian korelasi yang dilakukan terhadap 685 peubah prediktor memberikan informasi bahwa jumlah pengujian yang menghasilkan hubungan antar dua peubah prediktor lebih sedikit daripada tidak adanya hubungan antar dua peubah prediktor. Hasil uji korelasi antar peubah prediktor disajikan di Lampiran 3.

### 4.4. Penentuan Banyaknya Komponen

Langkah pertama dalam pemodelan PLSDA adalah menentukan banyaknya komponen yang terbentuk. Data gen pasien penderita kanker memiliki peubah prediktor yang cukup banyak sehingga metode PLS sangat cocok digunakan. Metode PLS akan mereduksi dimensi peubah awal menjadi beberapa komponen. Penentuan banyaknya komponen dilakukan dengan cara melihat proporsi keragaman dari peubah $X$ dan peubah $Y$ beserta kumulatifnya pada masing-masing komponen. Proporsi keragaman peubah $X$ dan peubah $Y$ beserta kumulatifnya dijelaskan di Lampiran 4. Pada Lampiran 4 dijelaskan banyaknya komponen yang terbentuk sebanyak 46 komponen tetapi tidak semua komponen akan dipilih. Banyaknya

komponen yang dipilih berdasarkan proporsi keragaman kumulatif dari peubah *X* dan peubah *Y* yang mencapai 80%.

Tabel 4.1. Proporsi Keragaman Peubah *X* dan Peubah *Y* beserta Kumulatif pada Komponen ke-41

| Komponen ke- | $R^2X$ | $R^2X$ Kumulatif | $R^2Y$ | $R^2Y$ Kumulatif |
|---|---|---|---|---|
| 41 | 0,0037 | 0,8018 | 0,0000 | 0,9819 |

Pada Tabel 4.1 dijelaskan bahwa pada komponen ke-41, $R^2X$ Kumulatif bernilai 80,18% dan $R^2Y$ Kumulatif bernilai 98,19%. Dapat disimpulkan bahwa sebanyak 41 komponen mampu menjelaskan keragaman dari peubah *X* dan peubah *Y* lebih dari 80%.

## 4.5. Pemodelan PLSDA

Setelah didapatkan banyaknya komponen yang digunakan, langkah berikutnya mengacak data kemudian data dibagi sesuai bentuk validasi silang yang digunakan. Bentuk validasi silang yang digunakan adalah *k-fold cross validation* dengan *k*=5. Data yang sudah diacak sebelumnya dibagi menjadi 5-*fold* dengan proporsi yang sama. Tiap *fold* terdiri dari 12 amatan. Tiap *fold* memiliki kesempatan satu kali menjadi data *testing* dan empat kali menjadi data *training*. Pengacakan dan pembagian *fold* dapat dilihat di Lampiran 5.

Setiap validasi silang terdiri dari 4 *fold* yang bertindak sebagai data *training* dan 1-*fold* yang bertindak sebagai data *testing*. Validasi silang dilakukan sebanyak 5 kali. Artinya, terdapat 5 model klasifikasi yang akan digunakan pada metode PLSDA. Untuk data *training* dibentuk model klasifikasi dengan metode PLSDA sedangkan data *testing* dilakukan untuk pengujian. Validasi silang dilakukan untuk mengetahui jumlah amatan yang tidak tepat diklasifikasikan sehingga akan diketahui akurasi pada masing-masing validasi silang. Dalam melakukan analisis diskriminan dengan metode PLSDA dibutuhkan bantuan *software* R Studio dengan *packages* DiscriMiner. Hasil validasi silang data *training* dan data *testing* masing-masing validasi dijelaskan pada Tabel 4.2 sampai Tabel 4.11.

Fungsi diskriminan pada PLSDA yang terbentuk sama dengan banyaknya kelompok pada peubah respon. Dalam penentuan kelompok pada peubah respon, digunakan metode skor fungsi

diskriminan yang tertinggi. Seluruh peubah prediktor dimasukkan ke dalam fungsi diskriminan yang terbentuk sehingga akan menghasilkan 5 skor diskriminan. Prediksi kelompok pada peubah respon dihasilkan dari skor diskriminan tertinggi pada suatu fungsi diskriminan.

### 4.5.1. Validasi Silang Pertama

Pada validasi silang pertama dilakukan pada *fold* 1 yang bertindak sebagai data *testing* sedangkan *fold* 2, *fold* 3, *fold* 4 dan *fold* 5 bertindak sebagai data *training*. Pada data *training* dibentuk model klasifikasi yang menghasilkan 5 fungsi diskriminan. Persamaan (4.1) merupakan fungsi diskriminan untuk peubah respon kelompok 1 (BRCA), persamaan (4.2) merupakan fungsi diskriminan untuk peubah respon kelompok 2 (COAD), persamaan (4.3) merupakan fungsi diskriminan untuk peubah respon kelompok 3 (KIRC), persamaan (4.4) merupakan fungsi diskriminan untuk peubah respon kelompok 4 (LUAD) dan persamaan (4.5) merupakan fungsi diskriminan untuk peubah respon kelompok 5 (PRAD).

$$D_1 = 0{,}9086 - 1{,}2456(10^{-3})X_1 + \cdots - 6{,}9629(10^{-4})X_{685} \quad (4.1)$$

$$D_2 = 0{,}0999 + 4{,}3673(10^{-4})X_1 + \cdots - 1{,}6600(10^{-3})X_{685} \quad (4.2)$$

$$D_3 = 0{,}8420 - 8{,}0902(10^{-4})X_1 + \cdots - 3{,}6698(10^{-3})X_{685} \quad (4.3)$$

$$D_4 = 0{,}2531 - 6{,}7544(10^{-4})X_1 + \cdots + 3{,}7428(10^{-3})X_{685} \quad (4.4)$$

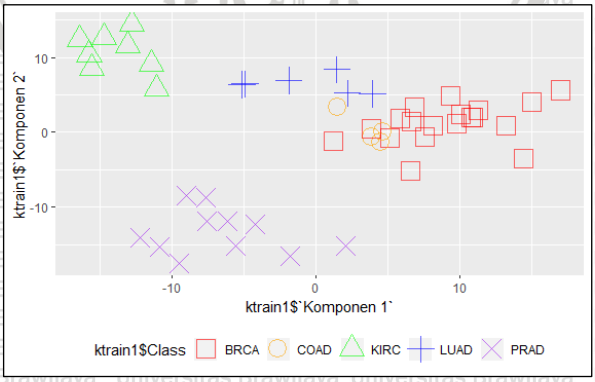
$$D_5 = -1{,}1036 + 9{,}4249(10^{-4})X_1 + \cdots + 2{,}2833(10^{-3})X_{685} \quad (4.5)$$

Data gen pasien sesuai amatan yang tergolong sebagai data *training* pada validasi silang pertama dimasukkan pada fungsi diskriminan pada persamaan (4.1) sampai persamaan (4.5). Prediksi seorang pasien berada di suatu kelompok kanker dihasilkan dari skor diskriminan tertinggi. Hasil klasifikasi data *training* dijelaskan pada Tabel 4.2.

Tabel 4.2. Tabel Kontingensi Data *Training* pada Validasi Silang Pertama

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 19 | 0 | 0 | 0 | 0 | 19 | 19 |
| 2 | 0 | 4 | 0 | 0 | 0 | 4 | 4 |
| 3 | 0 | 0 | 8 | 0 | 0 | 8 | 8 |
| 4 | 0 | 0 | 0 | 6 | 0 | 6 | 6 |
| 5 | 0 | 0 | 0 | 0 | 11 | 11 | 11 |
| Total | 19 | 4 | 8 | 6 | 11 | 48 | 48 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.2, pada validasi silang pertama yang dilakukan pada data *training* terdapat 48 dari 48 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang pertama, fungsi diskriminan yang terbentuk mampu memodelkan gen pasien ke dalam lima jenis kanker. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *training* pada validasi silang pertama ditunjukkan pada Gambar 4.2.




Gambar 4.2. *Scatter Plot* Data *Training* pada Validasi Silang Pertama

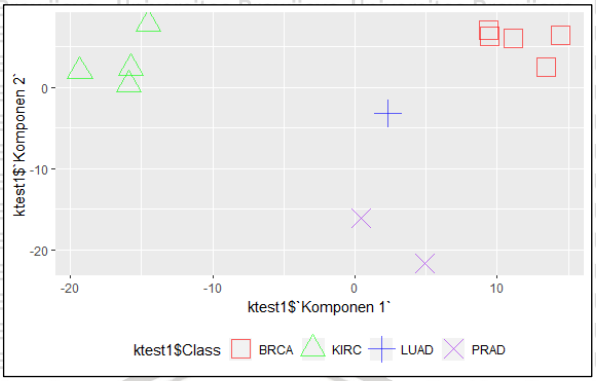
Gambar 4.2 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 48 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

Fungsi diskriminan yang terbentuk pada data *training*, kemudian diuji keakuratannya dalam memprediksi data baru dengan data yang bukan pembentuk model, yaitu data *testing*. Hasil klasifikasi data *testing* dijelaskan pada Tabel 4.3.

Tabel 4.3. Tabel Kontingensi Data *Testing* pada Validasi Silang Pertama

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 4 | 0 | 0 | 4 | 4 |
| 4 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 2 | 2 | 2 |
| Total | 5 | 0 | 4 | 1 | 2 | 12 | 12 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.3, pada validasi silang pertama yang dilakukan pada data *testing* terdapat 12 dari 12 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang pertama, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 100%. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *testing* pada validasi silang pertama ditunjukkan pada Gambar 4.3.




Gambar 4.3. *Scatter Plot* Data *Testing* pada Validasi Silang Pertama

Gambar 4.3 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 12 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

### 4.5.2. Validasi Silang Kedua

Pada validasi silang kedua dilakukan pada *fold* 2 yang bertindak sebagai data *testing* sedangkan *fold* 1, *fold* 3, *fold* 4 dan *fold* 5 bertindak sebagai data *training*. Pada data *training* dibentuk model klasifikasi yang menghasilkan 5 fungsi diskriminan. Persamaan (4.6) merupakan fungsi diskriminan untuk peubah respon kelompok 1 (BRCA), persamaan (4.7) merupakan fungsi diskriminan untuk peubah respon kelompok 2 (COAD), persamaan (4.8) merupakan fungsi diskriminan untuk peubah respon kelompok 3 (KIRC), persamaan (4.9) merupakan fungsi diskriminan untuk peubah respon kelompok 4 (LUAD) dan persamaan (4.10) merupakan fungsi diskriminan untuk peubah respon kelompok 5 (PRAD).

$$D_1 = 1{,}2093 - 2{,}1905(10^{-3})X_1 + \cdots - 1{,}2667(10^{-3})X_{685} \quad (4.6)$$

$$D_2 = -0{,}1231 + 7{,}3609(10^{-4})X_1 + \cdots - 1{,}2521(10^{-3})X_{685} \quad (4.7)$$

$$D_3 = 0{,}2065 - 2{,}0483(10^{-3})X_1 + \cdots - 3{,}0675(10^{-3})X_{685} \quad (4.8)$$

$$D_4 = 0{,}6418 + 9{,}5679(10^{-4})X_1 + \cdots + 3{,}0168(10^{-3})X_{685} \quad (4.9)$$

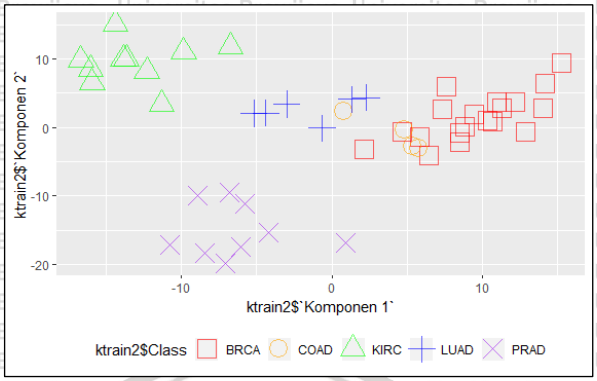
$$D_5 = -0{,}9345 + 2{,}5459(10^{-3})X_1 + \cdots + 2{,}5695(10^{-3})X_{685} \quad (4.10)$$

Data gen pasien sesuai amatan yang tergolong sebagai data *training* pada validasi silang kedua dimasukkan pada fungsi diskriminan pada persamaan (4.6) sampai persamaan (4.10). Prediksi seorang pasien berada di suatu kelompok kanker dihasilkan dari skor diskriminan tertinggi. Hasil klasifikasi data *training* dijelaskan pada Tabel 4.4.

Tabel 4.4. Tabel Kontingensi Data *Training* pada Validasi Silang Kedua

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 19 | 0 | 0 | 0 | 0 | 19 | 19 |
| 2 | 0 | 4 | 0 | 0 | 0 | 4 | 4 |
| 3 | 0 | 0 | 10 | 0 | 0 | 10 | 10 |
| 4 | 0 | 0 | 0 | 6 | 0 | 6 | 6 |
| 5 | 0 | 0 | 0 | 0 | 9 | 9 | 9 |
| Total | 19 | 4 | 10 | 6 | 9 | 48 | 48 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.4, pada validasi silang kedua yang dilakukan pada data *training* terdapat 48 dari 48 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang kedua, fungsi diskriminan yang terbentuk mampu memodelkan gen pasien ke dalam lima jenis kanker. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *training* pada validasi silang kedua ditunjukkan pada Gambar 4.4.




Gambar 4.4. *Scatter Plot* Data *Training* pada Validasi Silang Kedua

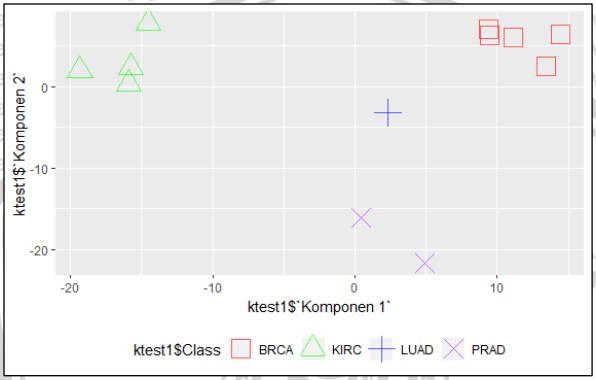
Gambar 4.4 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 48 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

Fungsi diskriminan yang terbentuk pada data *training*, kemudian diuji keakuratannya dalam memprediksi data baru dengan data yang bukan pembentuk model, yaitu data *testing*. Hasil klasifikasi data *testing* dijelaskan pada Tabel 4.5.

Tabel 4.5. Tabel Kontingensi Data *Testing* pada Validasi Silang Kedua

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 2 | 0 | 0 | 2 | 2 |
| 4 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 4 | 4 | 4 |
| Total | 5 | 0 | 2 | 1 | 4 | 12 | 12 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.5, pada validasi silang kedua yang dilakukan pada data *testing* terdapat 12 dari 12 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang kedua, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 100%. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *testing* pada validasi silang kedua ditunjukkan pada Gambar 4.5.




Gambar 4.5. *Scatter Plot* Data *Testing* pada Validasi Silang Kedua

Gambar 4.5 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 12 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

### 4.5.3. Validasi Silang Ketiga

Pada validasi silang ketiga dilakukan pada *fold* 3 yang bertindak sebagai data *testing* sedangkan *fold* 1, *fold* 2, *fold* 4 dan *fold* 5 bertindak sebagai data *training*. Pada data *training* dibentuk model klasifikasi yang menghasilkan 5 fungsi diskriminan. Persamaan (4.11) merupakan fungsi diskriminan untuk peubah respon kelompok 1 (BRCA), persamaan (4.12) merupakan fungsi diskriminan untuk 30

peubah respon kelompok 2 (COAD), persamaan (4.13) merupakan fungsi diskriminan untuk peubah respon kelompok 3 (KIRC), persamaan (4.14) merupakan fungsi diskriminan untuk peubah respon kelompok 4 (LUAD) dan persamaan (4.15) merupakan fungsi diskriminan untuk peubah respon kelompok 5 (PRAD).

$$D_1 = 0{,}6982 - 2{,}0920(10^{-4})X_1 + \cdots - 1{,}6826(10^{-3})X_{685} \quad (4.11)$$

$$D_2 = -0{,}0293 - 4{,}6197(10^{-4})X_1 + \cdots - 1{,}2070(10^{-3})X_{685} \quad (4.12)$$

$$D_3 = 1{,}0414 - 2{,}2648(10^{-3})X_1 + \cdots - 1{,}3171(10^{-3})X_{685} \quad (4.13)$$

$$D_4 = 0{,}0010 + 1{,}5771(10^{-3})X_1 + \cdots + 3{,}2793(10^{-3})X_{685} \quad (4.14)$$

$$D_5 = -0{,}7172 + 1{,}3588(10^{-3})X_1 + \cdots + 1{,}9274(10^{-3})X_{685} \quad (4.15)$$

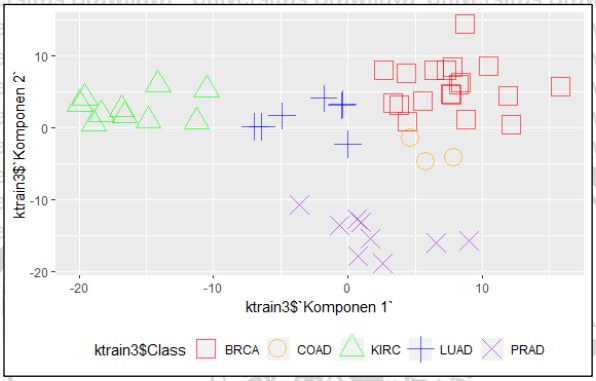
Data gen pasien sesuai amatan yang tergolong sebagai data *training* pada validasi silang ketiga dimasukkan pada fungsi diskriminan pada persamaan (4.11) sampai persamaan (4.15). Prediksi seorang pasien berada di suatu kelompok kanker dihasilkan dari skor diskriminan tertinggi. Hasil klasifikasi data *training* dijelaskan pada Tabel 4.6.

Tabel 4.6. Tabel Kontingensi Data *Training* pada Validasi Silang Ketiga

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 19 | 0 | 0 | 0 | 0 | 19 | 19 |
| 2 | 0 | 3 | 0 | 0 | 0 | 3 | 3 |
| 3 | 0 | 0 | 10 | 0 | 0 | 10 | 10 |
| 4 | 0 | 0 | 0 | 7 | 0 | 7 | 7 |
| 5 | 0 | 0 | 0 | 0 | 9 | 9 | 9 |
| Total | 19 | 3 | 10 | 7 | 9 | 48 | 48 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.6, pada validasi silang ketiga yang dilakukan pada data *training* terdapat 48 dari 48 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi.

Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang ketiga, fungsi diskriminan yang terbentuk mampu memodelkan gen pasien ke dalam lima jenis kanker. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *training* pada validasi silang ketiga ditunjukkan pada Gambar 4.6.




Gambar 4.6. *Scatter Plot* Data *Training* pada Validasi Silang Ketiga

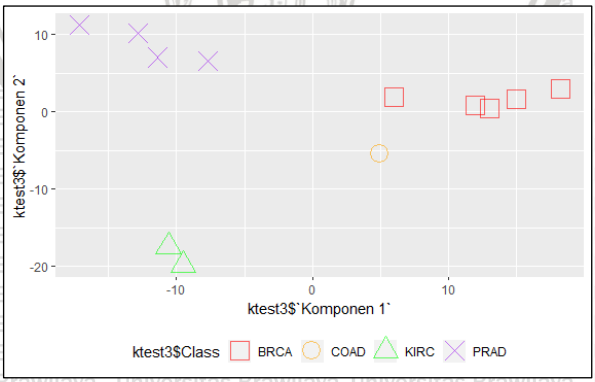
Gambar 4.6 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 48 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

Fungsi diskriminan yang terbentuk pada data *training*, kemudian diuji keakuratannya dalam memprediksi data baru dengan data yang bukan pembentuk model, yaitu data *testing*. Hasil klasifikasi data *testing* dijelaskan pada Tabel 4.7.

Tabel 4.7. Tabel Kontingensi Data *Testing* pada Validasi Silang Ketiga

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 2 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 2 | 0 | 0 | 2 | 2 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 4 | 4 | 4 |
| Total | 5 | 1 | 2 | 0 | 4 | 12 | 12 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.7, pada validasi silang ketiga yang dilakukan pada data *testing* terdapat 12 dari 12 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang ketiga, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 100%. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *testing* pada validasi silang ketiga ditunjukkan pada Gambar 4.7.




Gambar 4.7. *Scatter Plot* Data *Testing* pada Validasi Silang Ketiga

Gambar 4.7 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 12 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

### 4.5.4. Validasi Silang Keempat

Pada validasi silang keempat dilakukan pada *fold* 4 yang bertindak sebagai data *testing* sedangkan *fold* 1, *fold* 2, *fold* 3 dan *fold* 5 bertindak sebagai data *training*. Pada data *training* dibentuk model klasifikasi yang menghasilkan 5 fungsi diskriminan. Persamaan (4.16) merupakan fungsi diskriminan untuk peubah respon kelompok 1 (BRCA), persamaan (4.17) merupakan fungsi diskriminan untuk peubah respon kelompok 2 (COAD), persamaan (4.18) merupakan fungsi diskriminan untuk peubah respon kelompok 3 (KIRC), persamaan (4.19) merupakan fungsi diskriminan untuk peubah respon kelompok 4 (LUAD) dan persamaan (4.20) merupakan fungsi diskriminan untuk peubah respon kelompok 5 (PRAD).

$$D_1 = 0{,}2953 - 1{,}5729(10^{-3})X_1 + \cdots - 2{,}1644(10^{-3})X_{685} \quad (4.16)$$

$$D_2 = 0{,}0311 + 7{,}6051(10^{-4})X_1 + \cdots - 1{,}2611(10^{-3})X_{685} \quad (4.17)$$

$$D_3 = 0{,}7939 - 2{,}5693(10^{-3})X_1 + \cdots - 1{,}7147(10^{-3})X_{685} \quad (4.18)$$

$$D_4 = 0{,}0515 + 5{,}0333(10^{-4})X_1 + \cdots + 3{,}1358(10^{-3})X_{685} \quad (4.19)$$

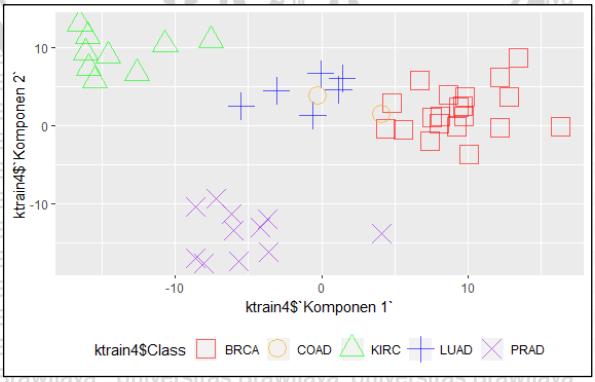
$$D_5 = -0{,}1718 + 2{,}8783(10^{-3})X_1 + \cdots + 2{,}0044(10^{-3})X_{685} \quad (4.20)$$

Data gen pasien sesuai amatan yang tergolong sebagai data *training* pada validasi silang keempat dimasukkan pada fungsi diskriminan pada persamaan (4.16) sampai persamaan (4.20). Prediksi seorang pasien berada di suatu kelompok kanker dihasilkan dari skor diskriminan tertinggi. Hasil klasifikasi data *training* dijelaskan pada Tabel 4.8.

Tabel 4.8. Tabel Kontingensi Data *Training* pada Validasi Silang Keempat

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 20 | 0 | 0 | 0 | 0 | 20 | 20 |
| 2 | 0 | 2 | 0 | 0 | 0 | 2 | 2 |
| 3 | 0 | 0 | 9 | 0 | 0 | 9 | 9 |
| 4 | 0 | 0 | 0 | 6 | 0 | 6 | 6 |
| 5 | 0 | 0 | 0 | 0 | 11 | 11 | 11 |
| Total | 20 | 2 | 9 | 6 | 11 | 48 | 48 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.8, pada validasi silang keempat yang dilakukan pada data *training* terdapat 48 dari 48 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang keempat, fungsi diskriminan yang terbentuk mampu memodelkan gen pasien ke dalam lima jenis kanker. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *training* pada validasi silang keempat ditunjukkan pada Gambar 4.8.




Gambar 4.8. *Scatter Plot* Data *Training* pada Validasi Silang Keempat

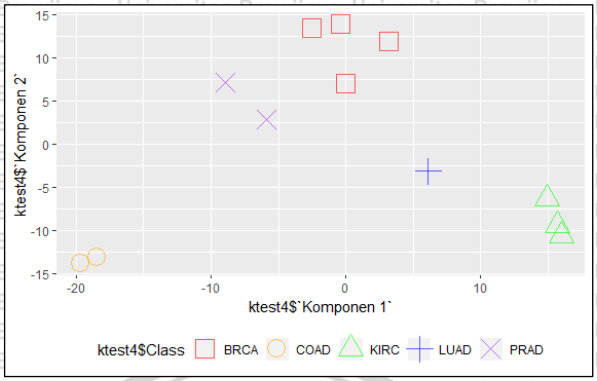
Gambar 4.8 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 48 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

Fungsi diskriminan yang terbentuk pada data *training*, kemudian diuji keakuratannya dalam memprediksi data baru dengan data yang bukan pembentuk model, yaitu data *testing*. Hasil klasifikasi data *testing* dijelaskan pada Tabel 4.9.

Tabel 4.9. Tabel Kontingensi Data *Testing* pada Validasi Silang Keempat

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 4 | 0 | 0 | 0 | 0 | 4 | 4 |
| 2 | 0 | 2 | 0 | 0 | 0 | 2 | 2 |
| 3 | 0 | 0 | 3 | 0 | 0 | 3 | 3 |
| 4 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 2 | 2 | 2 |
| Total | 4 | 2 | 3 | 1 | 2 | 12 | 12 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.8, pada validasi silang keempat yang dilakukan pada data *testing* terdapat 12 dari 12 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang keempat, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 100%. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *testing* pada validasi silang keempat ditunjukkan pada Gambar 4.9.




Gambar 4.9. *Scatter Plot* Data *Testing* pada Validasi Silang Keempat

Gambar 4.9 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 12 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

### 4.5.5. Validasi Silang Kelima

Pada validasi silang kelima dilakukan pada *fold* 5 yang bertindak sebagai data *testing* sedangkan *fold* 1, *fold* 2, *fold* 3 dan *fold* 4 bertindak sebagai data *training*. Pada data *training* dibentuk model klasifikasi yang menghasilkan 5 fungsi diskriminan. Persamaan (4.21) merupakan fungsi diskriminan untuk peubah respon kelompok 1 (BRCA), persamaan (4.22) merupakan fungsi diskriminan untuk peubah respon kelompok 2 (COAD), persamaan (4.23) merupakan fungsi diskriminan untuk peubah respon kelompok 3 (KIRC), persamaan (4.24) merupakan fungsi diskriminan untuk peubah respon kelompok 4 (LUAD) dan persamaan (4.25) merupakan fungsi diskriminan untuk peubah respon kelompok 5 (PRAD).

$$D_1 = 1{,}2446 - 2{,}0433(10^{-3})X_1 + \cdots - 7{,}1526(10^{-4})X_{685} \quad (4.21)$$

$$D_2 = -0{,}0860 - 1{,}6241(10^{-4})X_1 + \cdots - 2{,}0502(10^{-3})X_{685} \quad (4.22)$$

$$D_3 = 0{,}4500 - 2{,}8485(10^{-3})X_1 + \cdots - 2{,}6848(10^{-3})X_{685} \quad (4.23)$$

$$D_4 = -0{,}0004 + 2{,}1303(10^{-3})X_1 + \cdots + 2{,}4820(10^{-3})X_{685} \quad (4.24)$$

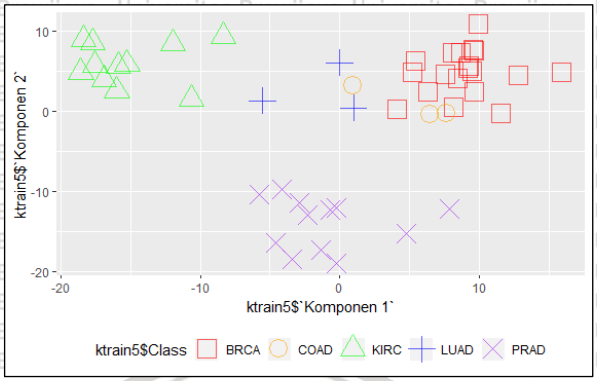
$$D_5 = -0{,}6082 + 2{,}9239(10^{-3})X_1 + \cdots + 2{,}9682(10^{-3})X_{685} \quad (4.25)$$

Data gen pasien sesuai amatan yang tergolong sebagai data *training* pada validasi silang kelima dimasukkan pada fungsi diskriminan pada persamaan (4.21) sampai persamaan (4.25). Prediksi seorang pasien berada di suatu kelompok kanker dihasilkan dari skor diskriminan tertinggi. Hasil klasifikasi data *training* dijelaskan pada Tabel 4.10.

Tabel 4.10. Tabel Kontingensi Data *Training* pada Validasi Silang Kelima

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 19 | 0 | 0 | 0 | 0 | 19 | 19 |
| 2 | 0 | 3 | 0 | 0 | 0 | 3 | 3 |
| 3 | 0 | 0 | 11 | 0 | 0 | 11 | 11 |
| 4 | 0 | 0 | 0 | 3 | 0 | 3 | 3 |
| 5 | 0 | 0 | 0 | 0 | 12 | 12 | 12 |
| Total | 19 | 3 | 11 | 3 | 12 | 48 | 48 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.10, pada validasi silang kelima yang dilakukan pada data *training* terdapat 48 dari 48 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang kelima, fungsi diskriminan yang terbentuk mampu memodelkan gen pasien ke dalam lima jenis kanker. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *training* pada validasi silang kelima ditunjukkan pada Gambar 4.10.




Gambar 4.10. *Scatter Plot* Data *Training* pada Validasi Silang Kelima

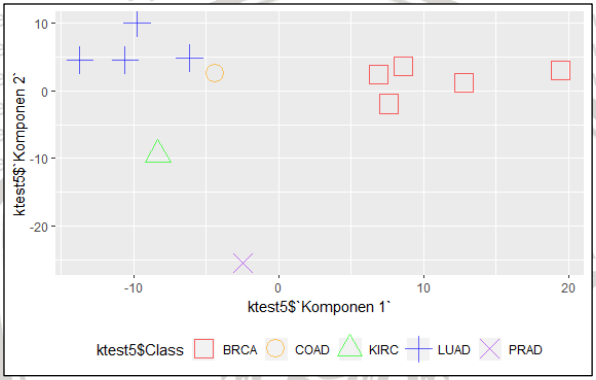
Gambar 4.10 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 48 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

Fungsi diskriminan yang terbentuk pada data *training*, kemudian diuji keakuratannya dalam memprediksi data baru dengan data yang bukan pembentuk model, yaitu data *testing*. Hasil klasifikasi data *testing* dijelaskan pada Tabel 4.11.

Tabel 4.11. Tabel Kontingensi Data *Testing* pada Validasi Silang Kelima

| Aktual | Prediksi | | | | | Total | Benar |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 2 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 4 | 3 | 0 | 0 | 4 | 0 | 4 | 4 |
| 5 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Total | 5 | 1 | 1 | 4 | 1 | 12 | 12 |
| Misklasifikasi | | | | | | 0 | |

Berdasarkan tabel kontingensi pada Tabel 4.11, pada validasi silang kelima yang dilakukan pada data *testing* terdapat 12 dari 12 amatan yang dideteksi benar dengan tidak terdapat misklasifikasi. Prediksi pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua benar. Dapat dikatakan bahwa pada validasi silang kelima, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 100%. Kemudian dibentuk *scatter plot* yang merupakan plot dari hasil dua komponen pertama yang telah didapatkan. *Scatter plot* data *testing* pada validasi silang kelima ditunjukkan pada Gambar 4.11.




Gambar 4.11. *Scatter Plot* Data *Testing* pada Validasi Silang Kelima

Gambar 4.11 menunjukkan bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah, sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Dapat disimpulkan bahwa sebanyak 12 pasien dengan pasien yang memiliki jenis kanker yang sama memiliki karakteristik yang sama dan pasien yang memiliki jenis kanker yang berbeda memiliki karakteristik yang berbeda.

## 4.6. Identifikasi Ketepatan Model Klasifikasi

Validasi silang yang telah dilakukan sebanyak lima kali, didapatkan nilai akurasi pada data *training* dan data *testing* pada masing-masing validasi silang. Tabel 4.12 merupakan nilai akurasi pada data *testing* dan data *training* masing-masing validasi silang secara keseluruhan.

Tabel 4.12. Nilai Akurasi Data *Testing* dan Data *Training*

| Validasi silang ke- | Akurasi Data *Testing* | Akurasi Data *Training* |
|---|---|---|
| 1 | 100% | 100% |
| 2 | 100% | 100% |
| 3 | 100% | 100% |
| 4 | 100% | 100% |
| 5 | 100% | 100% |
| Rata-rata | 100% | 100% |

Berdasarkan Tabel 4.12, dari kelima validasi silang didapatkan bahwa data *training* dan data *testing* memiliki rata-rata nilai akurasi yang sama yaitu sebesar 100%. Dapat disimpulkan bahwa semua model mampu mengklasifikasikan gen terhadap lima jenis kanker (BRCA, COAD, KIRC, LUAD, dan PRAD) dengan sempurna.

## 4.7. Fungsi Diskriminan

Pada validasi silang yang telah dilakukan sebelumnya didapatkan akurasi pada data *training* dan data *testing* sebesar 100%. Artinya, metode PLSDA sangat cocok digunakan untuk data gen pasien penderita kanker. Fungsi diskiriminan didapatkan dari keseluruhan data yang dimodelkan dengan PLSDA. Fungsi diskriminan yang terbentuk pada setiap kelompok dijelaskan pada persamaan (4.26) sampai persamaan (4.30).

$$D_1 = 1{,}0113 - 1{,}8782(10^{-2})X_1 + \cdots - 1{,}4114(10^{-3})X_{685} \quad (4.26)$$

$$D_2 = -0{,}0188 + 2{,}5702(10^{-4})X_1 + \cdots - 1{,}6679(10^{-3})X_{685} \quad (4.27)$$

$$D_3 = 0{,}5550 - 2{,}4277(10^{-3})X_1 + \cdots - 2{,}9665(10^{-3})X_{685} \quad (4.28)$$

$$D_4 = 0{,}0951 + 8{,}4341(10^{-4})X_1 + \cdots + 3{,}3923(10^{-3})X_{685} \quad (4.29)$$

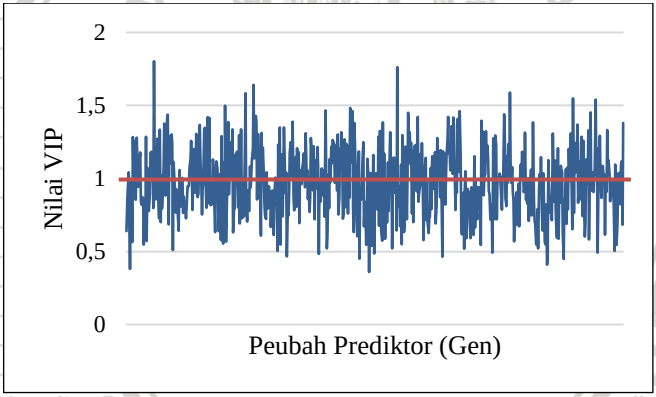
$$D_5 = -0{,}6427 + 2{,}4501(10^{-3})X_1 + \cdots + 2{,}6536(10^{-3})X_{685} \quad (4.30)$$

Persamaan (4.26) menjelaskan fungsi pengklasifikasian pada kanker jenis BRCA, persamaan (4.27) menjelaskan fungsi pengklasifikasian pada kanker jenis COAD, persamaan (4.28) menjelaskan fungsi pengklasifikasian pada kanker jenis KIRC, persamaan (4.29) menjelaskan fungsi pengklasifikasian pada kanker

jenis LUAD dan persamaan (4.30) menjelaskan pengklasifikasian pada kanker jenis PRAD. Fungsi pengklasifikasian yang telah dibentuk selanjutnya digunakan untuk membandingkan antara kategori Y awal dengan kategori Y prediksi. Kategori Y prediksi didapatkan dari skor diskriminan tertinggi. Hasil klasifikasi data dapat dilihat di Lampiran 6.

## 4.8. Identifikasi Peubah Penting dalam Model

Fungsi diskriminan yang telah terbentuk akan lebih baik diidentifikasi peubah prediktor mana yang berpengaruh besar terhadap pembentukan model. Dalam mengidentifikasi peubah penting dalam model digunakan nilai VIP (*Variable Importance in Projection*). Suatu peubah prediktor dikatakan memiliki peran dalam pembentukan model apabila nilai VIP > 1. Gambar 4.12 merupakan plot nilai VIP pada semua peubah prediktor. Nilai VIP masing-masing peubah prediktor dapat dilihat di Lampiran 7.

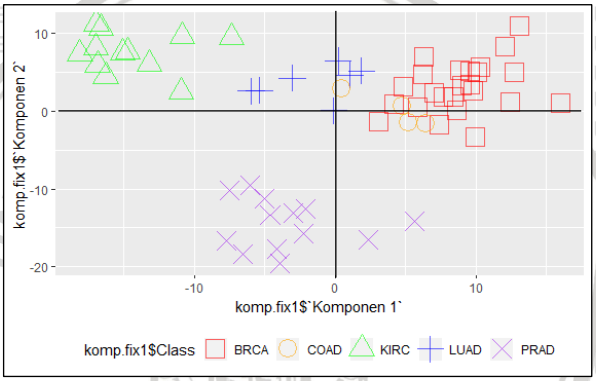



Gambar 4.12. Plot Nilai VIP

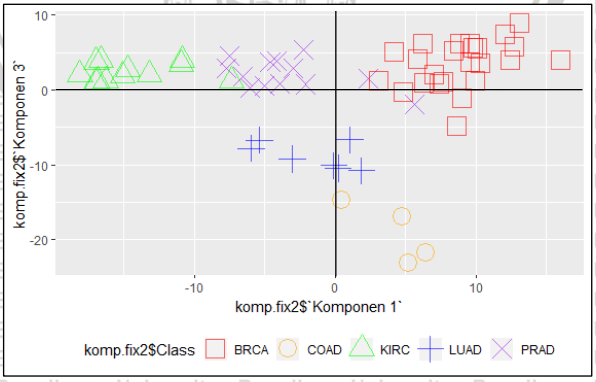
Pada Gambar 4.12 dapat dilihat bahwa sebagian besar gen 1 sampai gen 685 memiliki peranan yang penting dalam pembentukan model. Jumlah peubah prediktor sebanyak 685, terdapat 302 peubah prediktor yang memiliki nilai VIP > 1. Nilai VIP paling tinggi berada pada Gen 39 dengan nilai VIP sebesar 1,802749. Dapat disimpulkan bahwa Gen 39 merupakan peubah prediktor yang paling berpengaruh dalam pembentukan model.

## 4.9. Identifikasi Peubah Penciri Setiap Kategori Respon

Nilai VIP yang telah didapatkan pada sub bab sebelumnya kurang memberikan informasi mengenai peubah prediktor yang menjadi pembeda antara kategori satu dengan kategori lainnya. Untuk mengetahui peubah penciri pada masing-masing kategori respon dibutuhkan *scatter plot* dan *loading plot* yang dijelaskan pada Gambar 4.13 sampai dengan Gambar 4.17. *Scatter plot* dan *loading plot* didapatkan dari hasil pemodelan dengan metode PLSDA yang dilakukan pada keseluruhan data. Gambar 4.13 merupakan plot antara komponen 1 dengan komponen 2 dan Gambar 4.14 merupakan plot antara komponen 1 dengan komponen 3.

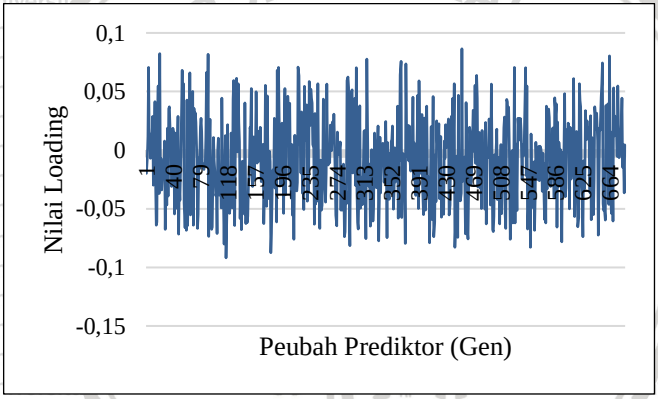



Gambar 4.13. *Scatter Plot* antara Komponen 1 dengan Komponen 2




Gambar 4.14. *Scatter Plot* antara Komponen 1 dengan Komponen 3

Berdasarkan Gambar 4.13 dan Gambar 4.14 dapat dilihat bahwa pasien penderita kanker yang sama berkumpul pada satu wilayah sedangkan pasien penderita kanker yang berbeda saling terpisah satu sama lain. Kanker jenis BRCA cenderung berada pada wilayah komponen 1 positif dan komponen 2 positif. Kanker jenis COAD cenderung berada pada wilayah komponen 1 positif dan komponen 2 negatif. Kanker jenis KIRC cenderung berada pada wilayah komponen 1 negatif dan komponen 3 positif. Kanker jenis LUAD cenderung berada pada wilayah komponen 1 positif dan komponen 3 negatif. Kanker jenis PRAD cenderung berada pada wilayah komponen 1 negatif dan komponen 2 negatif. Setelah diketahui wilayah dari masing-masing kategori respon, kemudian melihat puncak-puncak gen pada *loading plot* sesuai pada Gambar 4.15 sampai Gambar 4.17.

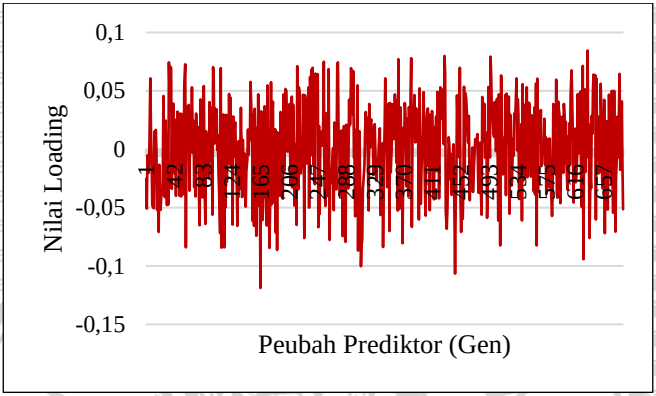



Gambar 4.15. *Loading Plot* pada Komponen 1

Berdasarkan *scatter plot* pada Gambar 4.13 dan Gambar 4.14, apabila jenis kanker berada di wilayah komponen 1 positif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling tinggi. Sebaliknya, apabila jenis kanker berada di wilayah komponen 1 negatif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling rendah. Tabel 4.13 menunjukkan nilai *loading* tertinggi dan terendah pada komponen 1.

Tabel 4.13. Nilai *Loading* pada Komponen 1

| | Peubah Prediktor | Nilai *Loading* |
|---|---|---|
| Tertinggi | gene_452 | 0,0865 |
| Terendah | gene_115 | -0,0919 |

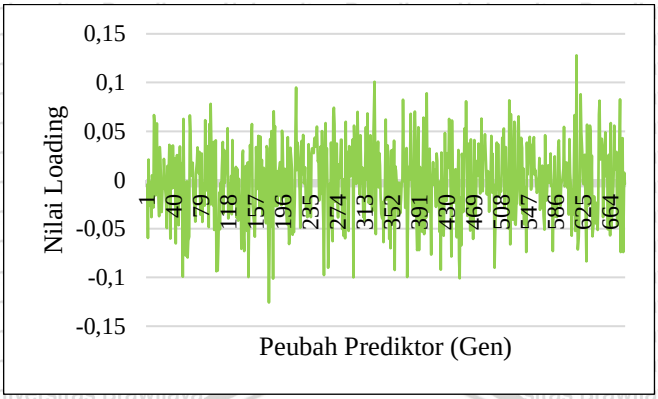



Gambar 4.16. *Loading Plot* pada Komponen 2

Berdasarkan *scatter plot* pada Gambar 4.13, apabila jenis kanker berada di wilayah komponen 2 positif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling tinggi. Sebaliknya, apabila jenis kanker berada di wilayah komponen 2 negatif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling rendah. Tabel 4.14 menunjukkan nilai *loading* tertinggi dan terendah pada komponen 2.

Tabel 4.14. Nilai *Loading* pada Komponen 2

| | Peubah Prediktor | Nilai *Loading* |
|---|---|---|
| Tertinggi | gene_634 | 0,0845 |
| Terendah | gene_165 | -0,1189 |




Gambar 4.17. *Loading Plot* pada Komponen 3

Berdasarkan *scatter plot* pada Gambar 4.14, apabila jenis kanker berada di wilayah komponen 3 positif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling tinggi. Sebaliknya, apabila jenis kanker berada di wilayah komponen 3 negatif, maka nilai *loading* yang dilihat adalah nilai *loading* yang paling rendah. Tabel 4.15 menunjukkan nilai *loading* tertinggi dan terendah pada komponen 3.

Tabel 4.15. Nilai *Loading* pada Komponen 3

| | Peubah Prediktor | Nilai *Loading* |
|---|---|---|
| Tertinggi | gene_616 | 0,1279 |
| Terendah | gene_176 | -0,1257 |

Nilai *loading* masing-masing komponen dapat dilihat di Lampiran 8. Berdasarkan Gambar 4.15 sampai Gambar 4.17 dapat dilihat bahwa Gen 452 dan Gen 634 merupakan peubah penciri pada kanker jenis BRCA, Gen 452 dan Gen 165 merupakan peubah penciri pada kanker jenis COAD, Gen 115 dan Gen 616 merupakan peubah penciri pada kanker jenis KIRC, Gen 452 dan Gen 176 merupakan peubah penciri pada kanker jenis LUAD serta Gen 115 dan Gen 165 merupakan peubah penciri pada kanker jenis PRAD.

# BAB V
# PENUTUP

## 5.1. Kesimpulan

Kesimpulan dari penelitian ini adalah:

1) Fungsi diskriminan yang terbentuk pada masing-masing jenis kanker seperti yang telah dijelaskan pada persamaan (4.26) sampai persamaan (4.30).

   Fungsi diskriminan untuk kanker jenis BRCA
   $D_1 = 1{,}0113 - 1{,}8782(10^{-2})X_1 + \cdots - 1{,}4114(10^{-3})X_{685}$

   Fungsi diskriminan untuk kanker jenis COAD
   $D_2 = -0{,}0188 + 2{,}5702(10^{-4})X_1 + \cdots - 1{,}6679(10^{-3})X_{685}$

   Fungsi diskriminan untuk kanker jenis KIRC
   $D_3 = 0{,}5550 - 2{,}4277(10^{-3})X_1 + \cdots - 2{,}9665(10^{-3})X_{685}$

   Fungsi diskriminan untuk kanker jenis LUAD
   $D_4 = -0{,}0951 + 8{,}4341(10^{-4})X_1 + \cdots + 3{,}3923(10^{-3})X_{685}$

   Fungsi diskriminan untuk kanker jenis PRAD
   $D_5 = -0{,}6427 + 2{,}4501(10^{-3})X_1 + \cdots + 2{,}6536(10^{-3})X_{685}$

   Fungsi diskriminan pada masing-masing jenis kanker yang telah terbentuk selanjutnya dapat digunakan untuk membandingkan antara kategori Y awal dengan kategori Y prediksi. Kategori Y prediksi didapatkan dari skor diskriminan yang paling tinggi.

2) Validasi silang yang dilakukan sebanyak lima kali terhadap data *training* dan data *testing* menghasilkan ketepatan klasifikasi sebesar 100%. Artinya tidak terdapat misklasifikasi pada kelima validasi silang.

3) Peubah penciri pada kanker jenis BRCA adalah Gen 452 dan Gen 634. Peubah penciri pada kanker jenis COAD adalah Gen 452 dan Gen 165. Peubah penciri pada kanker jenis KIRC adalah Gen 115 dan Gen 616. Peubah penciri pada kanker jenis LUAD adalah Gen 452 dan Gen 176. Peubah penciri pada kanker jenis PRAD adalah Gen 115 dan Gen 165.

## 5.2. Saran

Saran yang dapat diberikan untuk penelitian selanjutnya adalah sebelum dilakukan PLSDA sebaiknya data awal dilakukan *preprocessing*. Data yang cukup banyak dimungkinkan terdapat derau (*noise*) sehingga akan meningkatkan presisi kesimpulan yang didapat.

## DAFTAR PUSTAKA


Cinca, C. S. dan Nieto, B. G. 2011. *Partial Least Square Discriminant Analysis for Bankruptcy Prediction*. Spanyol: University of Zaragoza.

Enciso, M. P. dan Tenenhaus, M. 2003. *Prediction of Clinical Outcome with Microarray Data: A Partial Least Squares Discriminant Analysis (PLS-DA) Approach*. Paris: Springer-Verlag.

Farres, Platikanov, Tsakovski dan Tauler. 2015. Comparison of Variable Importance in Projection (VIP) and of the Selectivity Ratio (SR) Methods for Variable Selection and Interpretation. *Journal of Chemometrics*, Volume 29, pp. 528-536.

Hair, J. F. JR., Black, W. C., Babin, B. J. dan Anderson, R. E. 1998. *Multivariate Data Analysis*. Fifth Edition. New Jersey: Prentice Hall International Inc.

Hair, J. F. JR., Black, W. C., Babin, B. J. dan Anderson, R. E. 2010. *Multivariate Data Analysis*. Seventh Edition. New Jersey: Pearson Education Inc.

Han, J., Kamber, M. dan Pei, J. 2012. *Data Mining: Concepts and Techniques*. Third Edition. Morgan Kaufmann.

Huberty, J. C. 1934. *Applied MANOVA and Discriminant Analysis*. Second Edition. New York: John Willey and Sons.

Jaya, I G. N. M. dan Sumertajaya, I M. 2008. Pemodelan Persamaan Sruktural dengan Partial Least Square. *Jurnal Semnas Matematika dan Pendidikan Matematika*, Volume 1, pp. 118-132.

Johnson, R. A. dan Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*. Sixth Edition. New Jersey: Prentice Hall International Inc.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, Volume 2, pp. 1137-1145.

Mattjik, A. A. dan Sumertajaya, I M. 2011. *Sidik Peubah Ganda dengan Menggunakan SAS*. Bogor: IPB Press.

Meidawati, F. 2017. *Pemodelan Klasifikasi Obat Bahan Alam dengan Metode Partial Least Squares Discriminant Analysis*. Skripsi. Bogor: Institut Pertanian Bogor.

National Cancer Institute. 2018. The Cancer Genome Atlas. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers. Diakses pada 10 Maret 2019.

Ramzan, S. dan Khan, M. I. 2010. Dimension Reduction and Remedy of Multicollinearity Using Latent Variable Regression Methods. *World Applied Science Journal*, Volume 8, No. 4, 404–410.

Stefanska, Barbara dan MacEwan, David J. 2017. *Epigenetics and Gene Expression in Cancer, Inflammatory and Immune Disease*. New York: Springer Science + Business Media LLC.

Suryo. 1990. *Genetika Manusia*. Edisi Ketiga. Yogyakarta: Gadjah Mada University Press.

Walpole, R. E. 1993. *Pengantar Statistika*. Edisi Ketiga. Jakarta: PT Gramedia Pustaka Utama.

Wold, S., Sjostrom, M. dan Eriksson, L. 2001. PLS-Regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory System*, Volume 58, pp. 109–130.

# LAMPIRAN

Lampiran 1. Data Gen Pasien Penderita Kanker

| No | $X_1$ | $X_2$ | ... | $X_{684}$ | $X_{685}$ | $Y$ |
|---|---|---|---|---|---|---|
| 1 | 2,017209 | 3,265527 | ... | 6,748233 | 8,09189 | 5 |
| 2 | 0,592732 | 1,588421 | ... | 7,563318 | 5,709045 | 4 |
| 3 | 3,511759 | 4,327199 | ... | 5,438852 | 9,044247 | 5 |
| 4 | 3,663618 | 4,507649 | ... | 7,561143 | 7,511942 | 5 |
| 5 | 2,655741 | 2,821547 | ... | 7,763757 | 6,571285 | 1 |
| 6 | 3,467853 | 3,581918 | ... | 7,749903 | 9,18152 | 5 |
| 7 | 1,224966 | 1,691177 | ... | 8,406783 | 4,48542 | 3 |
| 8 | 2,854853 | 1,750478 | ... | 8,511859 | 8,484702 | 5 |
| 9 | 3,992125 | 2,77273 | ... | 8,551801 | 8,399398 | 1 |
| 10 | 3,642494 | 4,423558 | ... | 7,167789 | 8,102443 | 5 |
| 11 | 3,492071 | 3,553373 | ... | 6,493582 | 5,274243 | 1 |
| 12 | 2,941181 | 2,663276 | ... | 7,462878 | 4,865409 | 3 |
| 13 | 3,970348 | 2,364292 | ... | 6,861633 | 7,917575 | 5 |
| 14 | 1,551048 | 3,529846 | ... | 5,58526 | 7,14582 | 1 |
| 15 | 1,964842 | 2,18301 | ... | 6,189145 | 5,603448 | 1 |
| 16 | 2,901379 | 3,685368 | ... | 7,134786 | 7,307647 | 1 |
| 17 | 3,460913 | 3,618474 | ... | 7,959075 | 9,305647 | 4 |
| 18 | 3,004519 | 3,007178 | ... | 7,957781 | 5,069165 | 3 |
| 19 | 1,541465 | 2,54154 | ... | 8,310376 | 7,844982 | 3 |
| 20 | 4,167583 | 3,841389 | ... | 6,65695 | 7,42177 | 5 |
| 21 | 2,066916 | 2,619953 | ... | 6,724623 | 5,246913 | 1 |
| 22 | 3,529783 | 2,976712 | ... | 5,556751 | 7,354047 | 3 |
| 23 | 1,131853 | 2,351515 | ... | 6,724732 | 9,852687 | 4 |
| 24 | 3,121844 | 2,473943 | ... | 4,062536 | 4,464198 | 1 |
| 25 | 4,801852 | 2,648465 | ... | 7,72508 | 3,430205 | 3 |
| 26 | 4,317702 | 3,642678 | ... | 7,823093 | 8,669565 | 4 |

Lampiran 1. (Lanjutan)

| No | $X_1$ | $X_2$ | ... | $X_{684}$ | $X_{685}$ | $Y$ |
|---|---|---|---|---|---|---|
| 27 | 2,325242 | 3,247092 | ... | 7,370705 | 5,967355 | 2 |
| 28 | 0,657091 | 1,026304 | ... | 7,881395 | 3,41715 | 1 |
| 29 | 3,670081 | 3,382792 | ... | 7,463385 | 9,469176 | 1 |
| 30 | 5,688295 | 4,899949 | ... | 7,007184 | 10,97036 | 1 |
| 31 | 2,662479 | 2,742869 | ... | 6,50182 | 6,046229 | 1 |
| 32 | 2,090786 | 3,769116 | ... | 6,10415 | 6,702464 | 1 |
| 33 | 0 | 2,499578 | ... | 7,506034 | 7,209717 | 3 |
| 34 | 0 | 1,633152 | ... | 5,445078 | 7,087993 | 1 |
| 35 | 3,19022 | 5,690191 | ... | 7,382356 | 9,48823 | 5 |
| 36 | 2,464721 | 3,25408 | ... | 7,517929 | 6,08762 | 1 |
| 37 | 3,384243 | 2,673624 | ... | 6,984293 | 4,392956 | 3 |
| 38 | 4,416259 | 4,188978 | ... | 7,450931 | 9,096388 | 4 |
| 39 | 0 | 2,72275 | ... | 6,718677 | 3,567107 | 1 |
| 40 | 0 | 2,536525 | ... | 6,600484 | 4,593145 | 1 |
| 41 | 3,729379 | 2,293871 | ... | 7,870827 | 4,236577 | 3 |
| 42 | 3,026375 | 4,364103 | ... | 7,220997 | 6,834054 | 5 |
| 43 | 4,618861 | 5,648986 | ... | 8,344416 | 8,662658 | 5 |
| 44 | 3,757322 | 3,231048 | ... | 7,463524 | 5,474595 | 3 |
| 45 | 3,47241 | 3,007465 | ... | 6,574024 | 5,312462 | 3 |
| 46 | 3,092055 | 2,210327 | ... | 6,757343 | 7,419817 | 1 |
| 47 | 1,149259 | 2,575385 | ... | 7,75727 | 8,292069 | 5 |
| 48 | 3,44619 | 3,620962 | ... | 5,202625 | 3,392963 | 2 |
| 49 | 3,768449 | 2,736172 | ... | 7,337586 | 6,246425 | 1 |
| 50 | 4,577822 | 3,800237 | ... | 5,851732 | 10,45224 | 4 |
| 51 | 3,818472 | 3,215772 | ... | 6,731455 | 6,083198 | 1 |
| 52 | 3,858916 | 4,324523 | ... | 6,390575 | 6,952019 | 4 |
| 53 | 2,396708 | 2,399062 | ... | 5,463691 | 8,583534 | 1 |
| 54 | 3,639278 | 3,402176 | ... | 7,893162 | 7,403097 | 5 |

Lampiran 1. (Lanjutan)

| No | $X_1$ | $X_2$ | ... | $X_{684}$ | $X_{685}$ | $Y$ |
|---|---|---|---|---|---|---|
| 55 | 2,554196 | 1,296134 | ... | 7,308903 | 3,771695 | 2 |
| 56 | 2,350384 | 3,249369 | ... | 7,684089 | 8,389322 | 1 |
| 57 | 1,252355 | 2,40806 | ... | 5,787156 | 5,443172 | 1 |
| 58 | 3,667699 | 3,083179 | ... | 6,548558 | 6,923506 | 2 |
| 59 | 1,985792 | 3,640031 | ... | 7,138395 | 5,796351 | 3 |
| 60 | 1,870306 | 3,768777 | ... | 7,632646 | 6,189321 | 1 |

Keterangan:

$X$ : Panjang untaian RNA pada tingkat ekspresi gen (kbps)

$Y$ : Jenis kanker

1: BRCA
2: COAD
3: KIRC
4: LUAD
5: PRAD

Lampiran 2. *Missing Value* pada Setiap Peubah Prediktor

| Variable | Total Count | N | N Missing |
|---|---|---|---|
| gene_1 | 60 | 56 | 4 |
| gene_2 | 60 | 60 | 0 |
| gene_3 | 60 | 60 | 0 |
| gene_4 | 60 | 60 | 0 |
| gene_5 | 60 | 60 | 0 |
| gene_6 | 60 | 60 | 0 |
| gene_7 | 60 | 60 | 0 |
| gene_8 | 60 | 60 | 0 |
| gene_9 | 60 | 60 | 0 |
| gene_10 | 60 | 48 | 12 |
| gene_11 | 60 | 60 | 0 |
| gene_12 | 60 | 60 | 0 |
| gene_13 | 60 | 60 | 0 |
| gene_14 | 60 | 60 | 0 |
| gene_15 | 60 | 60 | 0 |
| gene_16 | 60 | 60 | 0 |
| gene_17 | 60 | 60 | 0 |
| gene_18 | 60 | 60 | 0 |
| gene_19 | 60 | 60 | 0 |
| gene_20 | 60 | 60 | 0 |
| gene_21 | 60 | 60 | 0 |
| gene_22 | 60 | 60 | 0 |
| gene_23 | 60 | 60 | 0 |
| gene_24 | 60 | 60 | 0 |
| gene_25 | 60 | 60 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| gene_684 | 60 | 60 | 0 |
| gene_685 | 60 | 60 | 0 |

Lampiran 3. Uji Korelasi

| | | $r$ | $t_{hit}$ | $t_{(0,025,58)}$ | $p$-$value$ | Hasil |
|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | 0,549 | 4,993 | 2,001717 | 0,000 | Signifikan |
| $X_1$ | $X_3$ | 0,164 | 1,269 | 2,001717 | 0,209 | Tidak Signifikan |
| $X_1$ | $X_4$ | -0,209 | -1,631 | 2,001717 | 0,108 | Tidak Signifikan |
| $X_1$ | $X_5$ | 0,161 | 1,239 | 2,001717 | 0,220 | Tidak Signifikan |
| $X_1$ | $X_6$ | -0,130 | -0,996 | 2,001717 | 0,323 | Tidak Signifikan |
| $X_1$ | $X_7$ | 0,106 | 0,810 | 2,001717 | 0,421 | Tidak Signifikan |
| $X_1$ | $X_8$ | 0,284 | 2,260 | 2,001717 | 0,028 | Signifikan |
| $X_1$ | $X_9$ | -0,355 | -2,891 | 2,001717 | 0,005 | Signifikan |
| $X_1$ | $X_{10}$ | 0,049 | 0,374 | 2,001717 | 0,710 | Tidak Signifikan |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $X_1$ | $X_{685}$ | 0,325 | 2,621 | 2,001717 | 0,011 | Signifikan |
| $X_2$ | $X_3$ | 0,018 | 0,134 | 2,001717 | 0,893 | Tidak signifikan |
| $X_2$ | $X_4$ | -0,069 | -0,530 | 2,001717 | 0,598 | Tidak signifikan |
| $X_2$ | $X_5$ | 0,304 | 2,426 | 2,001717 | 0,018 | Signifikan |
| $X_2$ | $X_6$ | 0,032 | 0,246 | 2,001717 | 0,806 | Tidak Signifikan |
| $X_2$ | $X_7$ | -0,210 | -1,635 | 2,001717 | 0,107 | Tidak Signifikan |
| $X_2$ | $X_8$ | 0,181 | 1,403 | 2,001717 | 0,166 | Tidak Signifikan |
| $X_2$ | $X_9$ | -0,195 | -1,512 | 2,001717 | 0,136 | Tidak Signifikan |
| $X_2$ | $X_{10}$ | 0,084 | 0,642 | 2,001717 | 0,524 | Tidak Signifikan |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $X_{681}$ | $X_{684}$ | -0,061 | -0,469 | 2,001717 | 0,641 | Tidak Signifikan |
| $X_{681}$ | $X_{685}$ | 0,236 | 1,852 | 2,001717 | 0,069 | Tidak Signifikan |
| $X_{682}$ | $X_{683}$ | -0,124 | -0,953 | 2,001717 | 0,344 | Tidak Signifikan |
| $X_{682}$ | $X_{684}$ | 0,115 | 0,878 | 2,001717 | 0,383 | Tidak Signifikan |
| $X_{682}$ | $X_{685}$ | 0,014 | 0,104 | 2,001717 | 0,918 | Tidak Signifikan |
| $X_{683}$ | $X_{684}$ | 0,043 | 0,327 | 2,001717 | 0,745 | Tidak Signifikan |
| $X_{683}$ | $X_{685}$ | -0,177 | -1,372 | 2,001717 | 0,175 | Tidak Signifikan |
| $X_{684}$ | $X_{685}$ | 0,100 | 0,768 | 2,001717 | 0,445 | Tidak Signifikan |

Lampiran 4. Proporsi Keragaman Peubah *X* dan Peubah *Y* beserta Kumulatif

| | R2X | R2Xcum | R2Y | R2Ycum |
|---|---|---|---|---|
| t1 | 0.133997994 | 0.1339980 | 2.516713e-01 | 0.2516713 |
| t2 | 0.110948524 | 0.2449465 | 2.289972e-01 | 0.4806685 |
| t3 | 0.073211168 | 0.3181577 | 2.033069e-01 | 0.6839754 |
| t4 | 0.042334489 | 0.3604922 | 1.875769e-01 | 0.8715523 |
| t5 | 0.071910939 | 0.4324031 | 3.806206e-02 | 0.9096144 |
| t6 | 0.037615824 | 0.4700189 | 3.940199e-02 | 0.9490164 |
| t7 | 0.043223240 | 0.5132422 | 1.475425e-02 | 0.9637706 |
| t8 | 0.025053618 | 0.5382958 | 9.743747e-03 | 0.9735144 |
| t9 | 0.025215085 | 0.5635109 | 4.215587e-03 | 0.9777300 |
| t10 | 0.013846819 | 0.5773577 | 2.834797e-03 | 0.9805648 |
| t11 | 0.010694633 | 0.5880523 | 6.600652e-04 | 0.9812248 |
| t12 | 0.010808775 | 0.5988611 | 3.213561e-04 | 0.9815462 |
| t13 | 0.010336329 | 0.6091974 | 1.678118e-04 | 0.9817140 |
| t14 | 0.009318609 | 0.6185160 | 6.222497e-05 | 0.9817762 |
| t15 | 0.008139794 | 0.6266558 | 4.948591e-05 | 0.9818257 |
| t16 | 0.008800066 | 0.6354559 | 2.036490e-05 | 0.9818461 |
| t17 | 0.010219803 | 0.6456757 | 1.312876e-05 | 0.9818592 |
| t18 | 0.007813451 | 0.6534892 | 1.001827e-05 | 0.9818692 |
| t19 | 0.007900844 | 0.6613900 | 4.257392e-06 | 0.9818735 |
| t20 | 0.007891428 | 0.6692814 | 3.403630e-06 | 0.9818769 |
| t21 | 0.006666989 | 0.6759484 | 2.467631e-06 | 0.9818794 |
| t22 | 0.008785990 | 0.6847344 | 1.702342e-06 | 0.9818811 |
| t23 | 0.007709789 | 0.6924442 | 9.378073e-07 | 0.9818820 |
| t24 | 0.008191069 | 0.7006353 | 6.516515e-07 | 0.9818827 |
| t25 | 0.007610332 | 0.7082456 | 5.367728e-07 | 0.9818832 |
| t26 | 0.007073884 | 0.7153195 | 3.776284e-07 | 0.9818836 |
| t27 | 0.007510196 | 0.7228297 | 1.566566e-07 | 0.9818837 |
| t28 | 0.005486209 | 0.7283159 | 1.858377e-07 | 0.9818839 |
| t29 | 0.006767314 | 0.7350832 | 9.643596e-08 | 0.9818840 |
| t30 | 0.006299942 | 0.7413831 | 6.427255e-08 | 0.9818841 |
| t31 | 0.008372785 | 0.7497559 | 1.763841e-08 | 0.9818841 |
| t32 | 0.005199526 | 0.7549555 | 1.948448e-08 | 0.9818841 |
| t33 | 0.006570431 | 0.7615259 | 1.154569e-08 | 0.9818841 |
| t34 | 0.005089281 | 0.7666152 | 7.938188e-09 | 0.9818841 |
| t35 | 0.005153439 | 0.7717686 | 2.557132e-09 | 0.9818841 |
| t36 | 0.004874749 | 0.7766434 | 1.567807e-09 | 0.9818841 |
| t37 | 0.005976285 | 0.7826196 | 1.080298e-09 | 0.9818841 |

Lampiran 4. (Lanjutan)

| | | | | |
|---|---|---|---|---|
| t38 | 0.004875882 | 0.7874955 | 5.601927e-10 | 0.9818841 |
| t39 | 0.004913357 | 0.7924089 | 2.956874e-10 | 0.9818841 |
| t40 | 0.005704361 | 0.7981132 | 9.105866e-11 | 0.9818841 |
| t41 | 0.003655271 | 0.8017685 | 1.036072e-10 | 0.9818841 |
| t42 | 0.004591537 | 0.8063601 | 1.653335e-11 | 0.9818841 |
| t43 | 0.005375854 | 0.8117359 | 6.813282e-12 | 0.9818841 |
| t44 | 0.005022680 | 0.8167586 | 3.636230e-12 | 0.9818841 |
| t45 | 0.005128321 | 0.8218869 | 1.491793e-12 | 0.9818841 |
| t46 | 0.004018803 | 0.8259057 | 1.924058e-13 | 0.9818841 |

Lampiran 5. Pembagian Data *Training* dan Data *Testing*

| Indeks | Amatan | *Fold* | CV 1 | CV 2 | CV 3 | CV 4 | CV 5 |
|---|---|---|---|---|---|---|---|
| 1 | 31 | 1 | | | | | |
| 2 | 19 | | | | | | |
| 3 | 25 | | | | | | |
| 4 | 40 | | | | | | |
| 5 | 5 | | | | | | |
| 6 | 13 | | | | | | |
| 7 | 15 | | | | | | |
| 8 | 54 | | | | | | |
| 9 | 33 | | | | | | |
| 10 | 22 | | | | | | |
| 11 | 52 | | | | | | |
| 12 | 28 | | | | | | |
| 13 | 6 | 2 | | | | | |
| 14 | 29 | | | | | | |
| 15 | 17 | | | | | | |
| 16 | 20 | | | | | | |
| 17 | 3 | | | | | | |
| 18 | 12 | | | | | | |
| 19 | 46 | | | | | | |
| 20 | 35 | | | | | | |
| 21 | 41 | | | | | | |
| 22 | 24 | | | | | | |
| 23 | 30 | | | | | | |
| 24 | 14 | | | | | | |

Keterangan:
: Data *Testing*
: Data *Training*

Lampiran 5. (Lanjutan)

| Indeks | Amatan | *Fold* | CV 1 | CV 2 | CV 3 | CV 4 | CV 5 |
|---|---|---|---|---|---|---|---|
| 25 | 53 | 3 | | | | | |
| 26 | 58 | | | | | | |
| 27 | 47 | | | | | | |
| 28 | 8 | | | | | | |
| 29 | 57 | | | | | | |
| 30 | 43 | | | | | | |
| 31 | 42 | | | | | | |
| 32 | 44 | | | | | | |
| 33 | 56 | | | | | | |
| 34 | 32 | | | | | | |
| 35 | 11 | | | | | | |
| 36 | 59 | | | | | | |
| 37 | 45 | 4 | | | | | |
| 38 | 51 | | | | | | |
| 39 | 16 | | | | | | |
| 40 | 26 | | | | | | |
| 41 | 48 | | | | | | |
| 42 | 49 | | | | | | |
| 43 | 1 | | | | | | |
| 44 | 55 | | | | | | |
| 45 | 4 | | | | | | |
| 46 | 34 | | | | | | |
| 47 | 18 | | | | | | |
| 48 | 7 | | | | | | |

Keterangan:

: Data *Testing*

: Data *Training*

Lampiran 5. (Lanjutan)

| Indeks | Amatan | *Fold* | CV 1 | CV 2 | CV 3 | CV 4 | CV 5 |
|---|---|---|---|---|---|---|---|
| 49 | 2 | 5 | | | | | |
| 50 | 9 | | | | | | |
| 51 | 23 | | | | | | |
| 52 | 21 | | | | | | |
| 53 | 60 | | | | | | |
| 54 | 10 | | | | | | |
| 55 | 50 | | | | | | |
| 56 | 36 | | | | | | |
| 57 | 39 | | | | | | |
| 58 | 38 | | | | | | |
| 59 | 37 | | | | | | |
| 60 | 27 | | | | | | |

Keterangan:
: Data *Testing*
: Data *Training*

Lampiran 6. Hasil Klasifikasi Data

| No | Y awal | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | Y prediksi |
|---|---|---|---|---|---|---|---|
| 1 | 5 | -0,0000224 | -0,0000663 | -0,0000940 | 0,0001253 | 1,0000570 | 5 |
| 2 | 4 | 0,0000569 | 0,0000093 | 0,0000230 | 0,9999805 | -0,0000697 | 4 |
| 3 | 5 | 0,0000689 | 0,0000108 | -0,0000085 | 0,0000322 | 0,9998966 | 5 |
| 4 | 5 | 0,0000983 | -0,0000373 | 0,0000263 | -0,0000949 | 1,0000070 | 5 |
| 5 | 1 | 0,9999843 | 0,0001056 | -0,0000533 | -0,0000308 | -0,0000058 | 1 |
| 6 | 5 | 0,0000567 | -0,0001228 | 0,0000896 | -0,0000226 | 0,9999991 | 5 |
| 7 | 3 | -0,0003973 | -0,0000053 | 1,0002420 | -0,0000726 | 0,0002333 | 3 |
| 8 | 5 | 0,0000395 | 0,0000966 | -0,0000723 | 0,0000112 | 0,9999249 | 5 |
| 9 | 1 | 0,9997924 | -0,0000864 | 0,0002408 | -0,0001451 | 0,0001983 | 1 |
| 10 | 5 | 0,0000289 | -0,0000352 | -0,0001599 | 0,0000019 | 1,0001640 | 5 |
| 11 | 1 | 0,9998250 | -0,0000617 | 0,0001774 | -0,0000133 | 0,0000727 | 1 |
| 12 | 3 | 0,0004743 | 0,0000179 | 0,9996677 | -0,0000013 | -0,0001586 | 3 |
| 13 | 5 | -0,0002822 | -0,0000393 | 0,0001698 | -0,0000916 | 1,0002430 | 5 |
| 14 | 1 | 0,9999042 | 0,0000567 | -0,0000817 | 0,0001115 | 0,0000092 | 1 |
| 15 | 1 | 0,9998553 | 0,0000390 | 0,0002103 | -0,0001370 | 0,0000324 | 1 |
| 16 | 1 | 1,0001510 | -0,0000343 | 0,0000031 | 0,0001342 | -0,0002540 | 1 |

Lampiran 6. (Lanjutan)

| No | Y awal | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | Y prediksi |
|---|---|---|---|---|---|---|---|
| 17 | 4 | -0,0001214 | -0,0000165 | -0,0000134 | 1,0000360 | 0,0001157 | 4 |
| 18 | 3 | -0,0000403 | 0,0001489 | 1,0000950 | -0,0001428 | -0,0000607 | 3 |
| 19 | 3 | 0,0001934 | -0,0001213 | 0,9999969 | -0,0001975 | 0,0001284 | 3 |
| 20 | 5 | 0,0004715 | 0,0001499 | -0,0004164 | 0,0001834 | 0,9996117 | 5 |
| 21 | 1 | 1,0001700 | -0,0000212 | -0,0001239 | -0,0000005 | -0,0000244 | 1 |
| 22 | 3 | -0,0000403 | -0,0000670 | 1,0000510 | 0,0000758 | -0,0000195 | 3 |
| 23 | 4 | -0,0000964 | 0,0000457 | 0,0001556 | 1,0000020 | -0,0001066 | 4 |
| 24 | 1 | 0,9998599 | -0,0000730 | 0,0000666 | 0,0000012 | 0,0001453 | 1 |
| 25 | 3 | -0,0001219 | 0,0000616 | 0,9998234 | -0,0000272 | 0,0002641 | 3 |
| 26 | 4 | 0,0002722 | -0,0000561 | -0,0002958 | 1,0001650 | -0,0000853 | 4 |
| 27 | 2 | 0,0000143 | 0,9999634 | 0,0000347 | -0,0000038 | -0,0000086 | 2 |
| 28 | 1 | 1,0001190 | 0,0000130 | -0,0000997 | 0,0000170 | -0,0000492 | 1 |
| 29 | 1 | 0,9996995 | 0,0000396 | 0,0003030 | -0,0000587 | 0,0000166 | 1 |
| 30 | 1 | 1,0001470 | -0,0000305 | -0,0001817 | 0,0000731 | -0,0000082 | 1 |
| 31 | 1 | 0,9996713 | -0,0000022 | 0,0002441 | -0,0000561 | 0,0001429 | 1 |
| 32 | 1 | 0,9998292 | -0,0000107 | 0,0000182 | 0,0000982 | 0,0000651 | 1 |

Lampiran 6. (Lanjutan)

| No | Y awal | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | Y prediksi |
|---|---|---|---|---|---|---|---|
| 33 | 3 | 0,0001584 | 0,0002449 | 0,9997992 | 0,0000816 | -0,0002842 | 3 |
| 34 | 1 | 1,0001200 | 0,0000109 | -0,0001814 | -0,0001277 | 0,0001779 | 1 |
| 35 | 5 | 0,0001369 | -0,0000022 | -0,0001165 | 0,0000266 | 0,9999552 | 5 |
| 36 | 1 | 0,9998888 | -0,0000905 | 0,0001091 | 0,0000400 | 0,0000526 | 1 |
| 37 | 3 | 0,0001981 | -0,0000632 | 0,9998728 | 0,0001528 | -0,0001605 | 3 |
| 38 | 4 | 0,0000979 | -0,0000220 | -0,0001488 | 1,0000020 | 0,0000710 | 4 |
| 39 | 1 | 1,0001050 | 0,0000359 | -0,0001768 | -0,0000014 | 0,0000371 | 1 |
| 40 | 1 | 1,0000140 | -0,0000729 | 0,0000709 | 0,0000269 | -0,0000391 | 1 |
| 41 | 3 | -0,0004755 | -0,0002084 | 1,0002340 | 0,0000099 | 0,0004400 | 3 |
| 42 | 5 | -0,0003266 | -0,0000239 | 0,0002736 | -0,0000073 | 1,0000840 | 5 |
| 43 | 5 | -0,0004678 | 0,0000513 | 0,0003448 | -0,0001023 | 1,0001740 | 5 |
| 44 | 3 | -0,0000682 | -0,0000358 | 1,0001760 | 0,0000634 | -0,0001350 | 3 |
| 45 | 3 | -0,0000978 | -0,0000193 | 1,0002210 | 0,0002227 | -0,0003264 | 3 |
| 46 | 1 | 1,0000610 | -0,0000540 | -0,0001335 | 0,0001937 | -0,0000674 | 1 |
| 47 | 5 | 0,0002028 | -0,0000702 | -0,0000698 | 0,0000556 | 0,9998816 | 5 |
| 48 | 2 | -0,0000022 | 0,9999498 | -0,0000491 | 0,0000984 | 0,0000030 | 2 |

Lampiran 6. (Lanjutan)

| No | Y awal | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | Y prediksi |
|---|---|---|---|---|---|---|---|
| 49 | 1 | 0,9999362 | 0,0000511 | 0,0000139 | -0,0000056 | 0,0000044 | 1 |
| 50 | 4 | 0,0001330 | 0,0000190 | 0,0001177 | 0,9998403 | -0,0001100 | 4 |
| 51 | 1 | 1,0001610 | -0,0000202 | -0,0000159 | -0,0000594 | -0,0000651 | 1 |
| 52 | 4 | -0,0003423 | 0,0000206 | 0,0001620 | 0,9999744 | 0,0001853 | 4 |
| 53 | 1 | 1,0002480 | -0,0000015 | -0,0002867 | 0,0000397 | 0,0000004 | 1 |
| 54 | 5 | -0,0000034 | 0,0000890 | 0,0000327 | -0,0001172 | 0,9999990 | 5 |
| 55 | 2 | 0,0000363 | 1,0000420 | -0,0000426 | -0,0000642 | 0,0000282 | 2 |
| 56 | 1 | 1,0001330 | 0,0000260 | -0,0000819 | 0,0000348 | -0,0001124 | 1 |
| 57 | 1 | 1,0002420 | 0,0000380 | -0,0000939 | -0,0001305 | -0,0000561 | 1 |
| 58 | 2 | -0,0000486 | 1,0000440 | 0,0000571 | -0,0000303 | -0,0000224 | 2 |
| 59 | 3 | 0,0002187 | 0,0000470 | 0,9998203 | -0,0001644 | 0,0000784 | 3 |
| 60 | 1 | 1,0000790 | 0,0001432 | 0,0000545 | -0,0000045 | -0,0002720 | 1 |

Lampiran 7. Nilai VIP

| Peubah Prediktor | Nilai VIP |
|---|---|
| gene_1 | 0,639771 |
| gene_2 | 0,807289 |
| gene_3 | 0,923852 |
| gene_4 | 1,042414 |
| gene_5 | 0,675748 |
| gene_6 | 0,380579 |
| gene_7 | 0,955355 |
| gene_8 | 0,978964 |
| gene_9 | 0,564037 |
| gene_10 | 1,283292 |
| gene_11 | 0,913456 |
| gene_12 | 1,252871 |
| gene_13 | 1,041569 |
| gene_14 | 0,854668 |
| gene_15 | 1,27821 |
| gene_16 | 1,132816 |
| gene_17 | 0,941793 |
| gene_18 | 0,994215 |
| gene_19 | 1,138955 |
| gene_20 | 1,179048 |
| gene_21 | 0,842813 |
| gene_22 | 0,818625 |
| gene_23 | 0,872645 |
| gene_24 | 0,864585 |
| gene_25 | 0,546174 |
| ⋮ | ⋮ |
| gene_684 | 0,683085 |
| gene_685 | 1,380432 |

Lampiran 8. Nilai *Loading*

| Komponen 1 | |
|---|---|
| gene_1 | -0,0197 |
| gene_2 | -0,0071 |
| gene_3 | 0,0117 |
| gene_4 | 0,0706 |
| gene_5 | 0,0151 |
| gene_6 | 0,0135 |
| gene_7 | -0,0069 |
| gene_8 | 0,0012 |
| gene_9 | 0,0131 |
| gene_10 | 0,0287 |
| ⋮ | ⋮ |
| gene_676 | 0,0220 |
| gene_677 | -0,0064 |
| gene_678 | 0,0035 |
| gene_679 | 0,0055 |
| gene_680 | 0,0215 |
| gene_681 | 0,0443 |
| gene_682 | -0,0147 |
| gene_683 | 0,0048 |
| gene_684 | -0,0364 |
| gene_685 | 0,0044 |

| Komponen 2 | |
|---|---|
| gene_1 | -0,0254 |
| gene_2 | -0,0508 |
| gene_3 | -0,0053 |
| gene_4 | -0,0115 |
| gene_5 | -0,0203 |
| gene_6 | 0,0020 |
| gene_7 | 0,0608 |
| gene_8 | 0,0138 |
| gene_9 | -0,0061 |
| gene_10 | -0,0478 |
| ⋮ | ⋮ |
| gene_676 | -0,0099 |
| gene_677 | 0,0279 |
| gene_678 | -0,0071 |
| gene_679 | 0,0319 |
| gene_680 | 0,0646 |
| gene_681 | -0,0177 |
| gene_682 | -0,0094 |
| gene_683 | 0,0409 |
| gene_684 | -0,0138 |
| gene_685 | -0,0515 |

| Komponen 3 | |
|---|---|
| gene_1 | -0,0214 |
| gene_2 | -0,0045 |
| gene_3 | -0,0595 |
| gene_4 | 0,0210 |
| gene_5 | -0,0263 |
| gene_6 | -0,0129 |
| gene_7 | -0,0019 |
| gene_8 | -0,0382 |
| gene_9 | 0,0050 |
| gene_10 | -0,0088 |
| ⋮ | ⋮ |
| gene_676 | 0,0271 |
| gene_677 | 0,0121 |
| gene_678 | 0,0829 |
| gene_679 | -0,0740 |
| gene_680 | 0,0325 |
| gene_681 | 0,0001 |
| gene_682 | 0,0429 |
| gene_683 | -0,0739 |
| gene_684 | 0,0076 |
| gene_685 | -0,0039 |

Lampiran 9. *Syntax* R Studio

```
data=read.csv(file.choose(),header=TRUE,
sep=";")
data

#KORELASI
korelasi=cor(data[,3:687])
write.csv(korelasi, file =
"D://Statistika/Skripsi/korelasigen.csv")

install.packages("DiscriMiner")
library(DiscriMiner)
library(plyr)
library(ggplot2)

#BANYAKNYA KOMPONEN
plsda1=plsDA(data[,3:687], data$Class,
autosel = TRUE)
plsda1
plsda1$R2

#PENGACAKAN DATA
set.seed(10)
fold=sample(60,60)
fold1=fold[1:12]
fold1
fold2=fold[13:24]
fold2
fold3=fold[25:36]
fold3
fold4=fold[37:48]
fold4
fold5=fold[49:60]
fold5

#CV 1
learn1=c(fold2, fold3, fold4, fold5)
test1=c(fold1)
```

Lampiran 9. (Lanjutan)

```
data1=data[learn1,]
data11=data[-learn1,]

##TRAIN 1
plsda1_train=plsDA(data1[,3:687],
data1$Class, autosel = FALSE, comps = 41)
plsda1_train$functions
plsda1_train$confusion
plsda1_train$error_rate

komp1.train=as.data.frame(plsda1_train$compo
nents)
komp11.train=as.matrix(komp1.train$'t1')
komp21.train=as.matrix(komp1.train$'t2')
komp.gab.train1=cbind(komp11.train,
komp21.train, data1$Class)
ktrain1=as.data.frame(komp.gab.train1)
colnames(ktrain1)=c("Komponen 1", "Komponen
2", "Class")
ktrain1[,3]=as.factor(ktrain1$Class)
ktrain1$Class=revalue(ktrain1$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktrain1, aes(x=ktrain1$`Komponen 1`,
y=ktrain1$`Komponen 2`)) +
  geom_point(aes(shape=ktrain1$Class,
color=ktrain1$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

##TEST 1
plsda1_test=plsDA(data11[,3:687],
data11$Class, autosel = FALSE, comps = 41)
plsda1_test$confusion
```

Lampiran 9. (Lanjutan)

```
plsda1_test$error_rate

komp1.test=as.data.frame(plsda1_test$compone
nts)
komp11.test=as.matrix(komp1.test$'t1')
komp21.test=as.matrix(komp1.test$'t2')
komp.gab.test1=cbind(komp11.test,
komp21.test, data11$Class)
ktest1=as.data.frame(komp.gab.test1)
colnames(ktest1)=c("Komponen 1", "Komponen
2", "Class")
ktest1[,3]=as.factor(ktest1$Class)
ktest1$Class=revalue(ktest1$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktest1, aes(x=ktest1$`Komponen 1`,
y=ktest1$`Komponen 2`)) +
  geom_point(aes(shape=ktest1$Class,
color=ktest1$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 2, 3, 4))
+
  scale_color_manual(values = c("red",
"green", "blue", "purple")) +
  theme(legend.position = "bottom")


#CV 2
learn2=c(fold1, fold3, fold4, fold5)
test2=c(fold2)
data2=data[learn2,]
data22=data[-learn2,]

##TRAIN 2
plsda2_train=plsDA(data2[,3:687],
data2$Class, autosel = FALSE, comps = 41)
plsda2_train$functions
plsda2_train$confusion
plsda2_train$error_rate
```

Lampiran 9. (Lanjutan)

```
komp2.train=as.data.frame(plsda2_train$compo
nents)
komp12.train=as.matrix(komp2.train$'t1')
komp22.train=as.matrix(komp2.train$'t2')
komp.gab.train2=cbind(komp12.train,
komp22.train, data2$Class)
ktrain2=as.data.frame(komp.gab.train2)
colnames(ktrain2)=c("Komponen 1", "Komponen
2", "Class")
ktrain2[,3]=as.factor(ktrain2$Class)
ktrain2$Class=revalue(ktrain2$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktrain2, aes(x=ktrain2$`Komponen 1`,
y=ktrain2$`Komponen 2`)) +
  geom_point(aes(shape=ktrain2$Class,
color=ktrain2$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

##TEST 2
plsda2_test=plsDA(data22[,3:687],
data22$Class, autosel = FALSE, comps = 41)
plsda2_test$confusion
plsda2_test$error_rate

komp2.test=as.data.frame(plsda2_test$compone
nts)
komp12.test=as.matrix(komp2.test$'t1')
komp22.test=as.matrix(komp2.test$'t2')
komp.gab.test2=cbind(komp12.test,
komp22.test, data22$Class)
ktest2=as.data.frame(komp.gab.test2)
```

Lampiran 9. (Lanjutan)

```
colnames(ktest2)=c("Komponen 1", "Komponen
2", "Class")
ktest2[,3]=as.factor(ktest2$Class)
ktest2$Class=revalue(ktest2$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktest1, aes(x=ktest1$`Komponen 1`,
y=ktest1$`Komponen 2`)) +
  geom_point(aes(shape=ktest1$Class,
color=ktest1$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 2, 3, 4))
+
  scale_color_manual(values = c("red",
"green", "blue", "purple")) +
  theme(legend.position = "bottom")

#CV 3
learn3=c(fold1, fold2, fold4, fold5)
test3=c(fold3)
data3=data[learn3,]
data33=data[-learn3,]

##TRAIN 3
plsda3_train=plsDA(data3[,3:687],
data3$Class, autosel = FALSE, comps = 41)
plsda3_train$functions
plsda3_train$confusion
plsda3_train$error_rate

komp3.train=as.data.frame(plsda3_train$compo
nents)
komp13.train=as.matrix(komp3.train$'t1')
komp23.train=as.matrix(komp3.train$'t2')
komp.gab.train3=cbind(komp13.train,
komp23.train, data3$Class)
ktrain3=as.data.frame(komp.gab.train3)
```

Lampiran 9. (Lanjutan)

```
colnames(ktrain3)=c("Komponen 1", "Komponen
2", "Class")
ktrain3[,3]=as.factor(ktrain3$Class)
ktrain3$Class=revalue(ktrain3$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktrain3, aes(x=ktrain3$`Komponen 1`,
y=ktrain3$`Komponen 2`)) +
  geom_point(aes(shape=ktrain3$Class,
color=ktrain3$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

##TEST 3
plsda3_test=plsDA(data33[,3:687],
data33$Class, autosel = FALSE, comps = 41)
plsda3_test$confusion
plsda3_test$error_rate

komp3.test=as.data.frame(plsda3_test$compone
nts)
komp13.test=as.matrix(komp3.test$'t1')
komp23.test=as.matrix(komp3.test$'t2')
komp.gab.test3=cbind(komp13.test,
komp23.test, data33$Class)
ktest3=as.data.frame(komp.gab.test3)
colnames(ktest3)=c("Komponen 1", "Komponen
2", "Class")
ktest3[,3]=as.factor(ktest3$Class)
ktest3$Class=revalue(ktest3$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktest3, aes(x=ktest3$`Komponen 1`,
y=ktest3$`Komponen 2`)) +
```

Lampiran 9. (Lanjutan)

```
  geom_point(aes(shape=ktest3$Class,
color=ktest3$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 4))
+
  scale_color_manual(values = c("red",
"orange", "green", "purple")) +
  theme(legend.position = "bottom")

#CV 4
learn4=c(fold1, fold2, fold3, fold5)
test4=c(fold4)
data4=data[learn4,]
data44=data[-learn4,]

##TRAIN 4
plsda4_train=plsDA(data4[,3:687],
data4$Class, autosel = FALSE, comps = 41)
plsda4_train$functions
plsda4_train$confusion
plsda4_train$error_rate

komp4.train=as.data.frame(plsda4_train$compo
nents)
komp14.train=as.matrix(komp4.train$'t1')
komp24.train=as.matrix(komp4.train$'t2')
komp.gab.train4=cbind(komp14.train,
komp24.train, data4$Class)
ktrain4=as.data.frame(komp.gab.train4)
colnames(ktrain4)=c("Komponen 1", "Komponen
2", "Class")
ktrain4[,3]=as.factor(ktrain4$Class)
ktrain4$Class=revalue(ktrain4$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktrain4, aes(x=ktrain4$`Komponen 1`,
y=ktrain4$`Komponen 2`)) +
```

Lampiran 9. (Lanjutan)

```
  geom_point(aes(shape=ktrain4$Class,
color=ktrain4$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

##TEST 4
plsda4_test=plsDA(data44[,3:687],
data44$Class, autosel = FALSE, comps = 41)
plsda4_test$confusion
plsda4_test$error_rate

komp4.test=as.data.frame(plsda4_test$compone
nts)
komp14.test=as.matrix(komp4.test$'t1')
komp24.test=as.matrix(komp4.test$'t2')
komp.gab.test4=cbind(komp14.test,
komp24.test, data44$Class)
ktest4=as.data.frame(komp.gab.test4)
colnames(ktest4)=c("Komponen 1", "Komponen
2", "Class")
ktest4[,3]=as.factor(ktest4$Class)
ktest4$Class=revalue(ktest4$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktest4, aes(x=ktest4$`Komponen 1`,
y=ktest4$`Komponen 2`)) +
  geom_point(aes(shape=ktest4$Class,
color=ktest4$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")
```

Lampiran 9. (Lanjutan)

```
#CV 5
learn5=c(fold1, fold2, fold3, fold4)
test5=c(fold5)
data5=data[learn5,]
data55=data[-learn5,]

##TRAIN 5
plsda5_train=plsDA(data5[,3:687],
data5$Class, autosel = FALSE, comps = 41)
plsda5_train$functions
plsda5_train$confusion
plsda5_train$error_rate

komp5.train=as.data.frame(plsda5_train$compo
nents)
komp15.train=as.matrix(komp5.train$'t1')
komp25.train=as.matrix(komp5.train$'t2')
komp.gab.train5=cbind(komp15.train,
komp25.train, data5$Class)
ktrain5=as.data.frame(komp.gab.train5)
colnames(ktrain5)=c("Komponen 1", "Komponen
2", "Class")
ktrain5[,3]=as.factor(ktrain5$Class)
ktrain5$Class=revalue(ktrain5$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktrain5, aes(x=ktrain5$`Komponen 1`,
y=ktrain5$`Komponen 2`)) +
  geom_point(aes(shape=ktrain5$Class,
color=ktrain5$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")
```

Lampiran 9. (Lanjutan)

```
##TEST 5
plsda5_test=plsDA(data55[,3:687],
data55$Class, autosel = FALSE, comps = 41)
plsda5_test$confusion
plsda5_test$error_rate

komp5.test=as.data.frame(plsda5_test$compone
nts)
komp15.test=as.matrix(komp5.test$'t1')
komp25.test=as.matrix(komp5.test$'t2')
komp.gab.test5=cbind(komp15.test,
komp25.test, data55$Class)
ktest5=as.data.frame(komp.gab.test5)
colnames(ktest5)=c("Komponen 1", "Komponen
2", "Class")
ktest5[,3]=as.factor(ktest5$Class)
ktest5$Class=revalue(ktest5$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(ktest5, aes(x=ktest5$`Komponen 1`,
y=ktest5$`Komponen 2`)) +
  geom_point(aes(shape=ktest5$Class,
color=ktest5$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

#PLSDA
plsda.fix=plsDA(data[,3:687], data$Class,
autosel = FALSE, comps = 41)
plsda.fix
plsda.fix$functions
plsda.fix$scores
plsda.fix$VIP
plsda.fix$loadings
```

Lampiran 9. (Lanjutan)

```
#MENGURUTKAN VIP
vip=as.data.frame(plsda.fix$VIP)
vip.fix=as.matrix(vip$`Model VIP`)
vipx=matrix(0, nrow(vip.fix), 1)
s=0
for (i in 1:nrow(vip.fix)) {
  vipx[i,]=s+i
  vipx[i,]=vipx[i,]
}
vipxx=cbind(vipx, vip.fix)
vip.urut=as.matrix(vipxx[order(vipxx[,2],
decreasing = T)])
vipxx.urut=cbind(vip.urut,
vip.fix[vip.urut])

#COMPONENT
komp=as.data.frame(plsda.fix$components)
komp1=as.matrix(komp$'t1')
komp2=as.matrix(komp$'t2')
komp.gab1=cbind(komp1, komp2, data$Class)
komp.fix1=as.data.frame(komp.gab1)
colnames(komp.fix1)=c("Komponen 1",
"Komponen 2", "Class")
komp.fix1[,3]=as.factor(komp.fix1$Class)
komp.fix1$Class=revalue(komp.fix1$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(komp.fix1, aes(x=komp.fix1$`Komponen
1`, y=komp.fix1$`Komponen 2`)) +
  geom_point(aes(shape=komp.fix1$Class,
color=komp.fix1$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")
```

Lampiran 9. (Lanjutan)

```
komp=as.data.frame(plsda.fix$components)
komp1=as.matrix(komp$'t1')
komp3=as.matrix(komp$'t3')
komp.gab2=cbind(komp1, komp3, data$Class)
komp.fix2=as.data.frame(komp.gab2)
colnames(komp.fix2)=c("Komponen 1",
"Komponen 3", "Class")
komp.fix2[,3]=as.factor(komp.fix2$Class)
komp.fix2$Class=revalue(komp.fix2$Class,
c("1"="BRCA", "2"="COAD", "3"="KIRC",
"4"="LUAD", "5"="PRAD"))
ggplot(komp.fix2, aes(x=komp.fix2$`Komponen
1`, y=komp.fix2$`Komponen 3`)) +
  geom_point(aes(shape=komp.fix2$Class,
color=komp.fix2$Class), size=6, alpha=0.6) +
  scale_shape_manual(values = c(0, 1, 2, 3,
4)) +
  scale_color_manual(values = c("red",
"orange", "green", "blue", "purple")) +
  theme(legend.position = "bottom")

#MENGURUTKAN LOADING
load1=as.data.frame(plsda.fix$loadings)
load1.fix=as.matrix(load1$`w*1`)
load1x=matrix(0, nrow(load1.fix), 1)
s=0
for (i in 1:nrow(load1.fix)) {
  load1x[i,]=s+i
  load1x[i,]=load1x[i,]
}
load1xx=cbind(load1x, load1.fix)

load2=as.data.frame(plsda.fix$loadings)
load2.fix=as.matrix(load2$`w*2`)
load2x=matrix(0, nrow(load2.fix), 1)
s=0
for (i in 1:nrow(load2.fix)) {
```

Lampiran 9. (Lanjutan)

```
  load2x[i,]=s+i
  load2x[i,]=load2x[i,]
}
load2xx=cbind(load2x, load2.fix)

load3=as.data.frame(plsda.fix$loadings)
load3.fix=as.matrix(load3$`w*3`)
load3x=matrix(0, nrow(load3.fix), 1)
s=0
for (i in 1:nrow(load3.fix)) {
  load3x[i,]=s+i
  load3x[i,]=load3x[i,]
}
load3xx=cbind(load3x, load3.fix)

##BRCA
load1.urut1=as.matrix(load1xx[order(load1xx[
,2], decreasing = T)])
load1xx.urut1=cbind(load1.urut1,
load1.fix[load1.urut1])

load2.urut1=as.matrix(load2xx[order(load2xx[
,2], decreasing = T)])
load2xx.urut1=cbind(load2.urut1,
load2.fix[load2.urut1])

##COAD
load1.urut2=as.matrix(load1xx[order(load1xx[
,2], decreasing = T)])
load1xx.urut2=cbind(load1.urut2,
load1.fix[load1.urut2])

load2.urut2=as.matrix(load2xx[order(load2xx[
,2], decreasing = F)])
load2xx.urut2=cbind(load2.urut2,
load2.fix[load2.urut2])
```

Lampiran 9. (Lanjutan)

```
##KIRC
load1.urut3=as.matrix(load1xx[order(load1xx[
,2], decreasing = F)])
load1xx.urut3=cbind(load1.urut3,
load1.fix[load1.urut3])

load3.urut3=as.matrix(load3xx[order(load3xx[
,2], decreasing = T)])
load3xx.urut3=cbind(load3.urut3,
load3.fix[load3.urut3])

##LUAD
load1.urut4=as.matrix(load1xx[order(load1xx[
,2], decreasing = T)])
load1xx.urut4=cbind(load1.urut4,
load1.fix[load1.urut4])

load3.urut4=as.matrix(load3xx[order(load3xx[
,2], decreasing = F)])
load3xx.urut4=cbind(load3.urut4,
load3.fix[load3.urut4])

##PRAD
load1.urut5=as.matrix(load1xx[order(load1xx[
,2], decreasing = F)])
load1xx.urut5=cbind(load1.urut5,
load1.fix[load1.urut5])

load2.urut5=as.matrix(load2xx[order(load2xx[
,2], decreasing = F)])
load2xx.urut5=cbind(load2.urut5,
load2.fix[load2.urut5])
```